

# Data Profiling

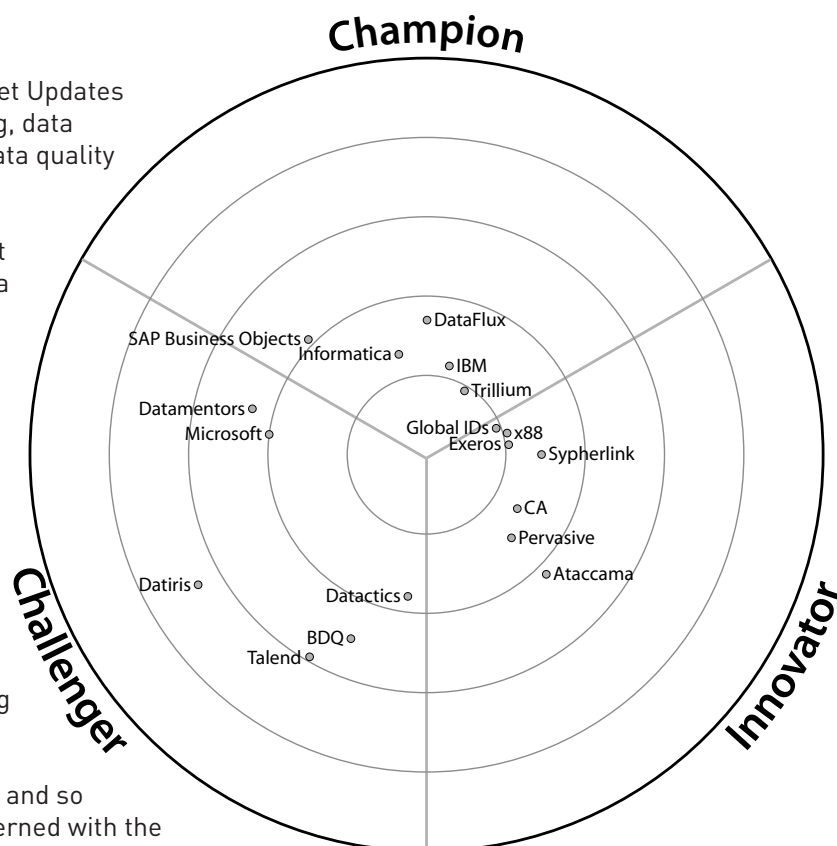
## Introduction

This is the second of four Market Updates on data discovery, data profiling, data cleansing and matching, and data quality platforms respectively.

For those readers who have not seen the Market Update on data discovery we need to explain the distinction between that discipline and data profiling. Data profiling does two things: it discovers relationships between data elements, whether they are in the same data source or across multiple, heterogeneous data sources; and it performs statistical analysis against individual columns (in a relational database) discovering such things as the number of null values, whether the data matches the expected datatype and so on. Data discovery is only concerned with the first of these capabilities. So why make the distinction? There are two answers: the first is that there are data discovery tools that are not data profiling tools and the second is that data profiling is closely associated with data cleansing whereas data discovery has utility in a number of other areas, for example it is complementary to data modelling. For a full discussion on this topic see the first Market Update in this series and its accompanying Spotlight Paper.

This is, in fact, a problem. This association has meant that a number of the leading vendors have failed to exploit the capabilities of their profiling tools as they might have done, also failing to introduce the sort of functionality that might be used to support non-data quality projects. Indeed, it is notable that we have different reports from the major suppliers as to the popularity of data profiling tools on a stand-alone basis. Some have reported that they have seen a drop-off in demand over the last couple of years while others have told us the complete opposite. We suspect that this exactly reflects the extent to which different vendors have appreciated the potential for using data profiling outside of data quality environments.

Leaving that aside this Market Update is specifically about data profiling: it includes consideration of those aspects of data profiling products that support data discovery in general terms but also pays particular



**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.

attention to the way in which data profiling provides facilities that complement data cleansing and matching.

## Key market issues

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 four ways.

Firstly, the extent to which the product supports multiple, heterogeneous data sources: a few products, such as the open source Talend Open Profiler, are limited to profiling a single data source at a time; most can handle a handful of sources, but only a few suppliers can handle large numbers of sources. This, if you like, is a measure of scalability (as is the ability to perform appropriately with very large tables that contain rows numbered in the billions—not all tools can handle this sort of scale). 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.

Secondly, there are the data discovery facilities provided. Since these are covered in depth in the papers mentioned previously we will not discuss these in detail but the sorts of facilities you would like to have include, but are not limited to, the identification of redundant columns, primary/foreign key discovery (even when these are of different datatypes or field lengths), business and transformation rule discovery, exception detection against discovered or pre-defined 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.

However, 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 data

rules) that have been pre-defined 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.

Thirdly, data profiling is, or can be, an important collaborative tool. It is typically business analysts and domain stewards 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 capture or use quality metrics.

Fourthly, 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. Now 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, most tools cannot cope with multiple rule violations. What you really like to know is what percentage of records have no violations, one violations, 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 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.

## Vendor landscape

There are more than 20 vendors involved in the data profiling space. Of these, 18 responded to our requests for information while Oracle and Human Inference declined to be involved, the former because it is in the throes of merging its Oracle Warehouse Builder and Oracle Data Integrator teams and Human Inference because of its strong focus on data quality rather than profiling, though it does offer the latter.

There are three camps within which data profiling vendors fall; those that only offer data profiling, those that focus exclusively on data quality, and those that offer broader sets of capabilities. In the first category are BDQ (though this company also offers a product aimed specifically at data stewards and governance), Datoris, Exeros, Sypherlink and x88; in the second are Datactics, Datamentors and Trillium; and in the third group are Ataccama, DataFlux, Global IDs, IBM, Informatica, Microsoft, Pervasive, SAP Business Objects and Talend, the last of these also being in a separate category in that it is an open source product.

Before proceeding further we should mention two recently released third party products. The first of these is PartyQualityInsight from DataQualityFirst. This is a tool that supplements IBM environments for validating business rules relating to parties (customers, suppliers, employees and so on) in CDI (customer data integration) projects. The second is Data Validator from DVO, which provides automated data testing and business rule checking (but not discovery) for Informatica environments. In our view Data Validator should be a no-brainer for Informatica customers as it removes a significant part of the manual testing that would otherwise be required, not only for data movement and data quality purposes but also when testing against upgrades of Informatica PowerCenter.

We should also note a number of partnerships. The most important of these is between CA and Exeros with the former now reselling Exeros X-Profiler as CA Data Profiler to work alongside its ERWin data modeling tool, with Exeros' more advanced capabilities being available from Exeros as an up-sell. In addition, both BDQ and Datactics are embedded in third party data quality platforms. The same is true of Ataccama whose technology is resold by iWay. iWay has embedded Ataccama's Data Quality Center (and Master Data Center) into its enterprise service bus. This means that you can profile

(and cleanse and enrich) data in real-time at much greater speeds and volumes than would normally be the case with other vendors. Sypherlink has a long-standing partnership with ASG (and that company's Rochade repository) and it also has a unique capability in that it generates ETL (extract, transform and load) transformations based on the relationships it discovers and can thus act as a pre-cursor to the use of ETL tools. While not a partnership, we should also note that Microsoft has yet to integrate Zoomix (which it acquired in 2008) into its existing capabilities, though it plans to do so.

## Summary and conclusions

We noted previously that the traditional data quality vendors have tended to ignore the potential that data profiling offers. At least partly for this reason most of the leading products in this Market Update, from a technical perspective, are offered by smaller vendors. However, such suppliers have obvious drawbacks such as limited geographic coverage, as a result of which many users will continue to prefer a big name provider. Of these, we believe Trillium is some way ahead of its major competitors in terms of data discovery though if you include DVO's Data Validator along with Informatica Data Explorer (which we have not) then this would significantly narrow the gap for Informatica. The other major vendors to seriously consider are IBM and CA, in the latter case thanks to its partnership with Exeros, though CA will be focused on data discovery to augment data modelling rather than for other purposes.

Of the smaller players we would single out Exeros, Global IDs, Sypherlink and x88 as the leading innovators in this market, along with Ataccama, especially when used in conjunction with iWay's Integration Server.

*Philip Howard  
Data Management  
February 2009*