

Data Discovery

Introduction

This is the first of four Market Updates on data discovery, data profiling, data cleansing and matching, and data quality platforms respectively. Since data discovery is a new market sector we need to make a distinction between it and data profiling. We define data discovery or, more correctly, data relationship discovery, as “the discovery of relationships between data elements, regardless of where the data is stored”. Data profiling tools do this but they also perform statistical analysis against data sources for such things as the number of null values that are specifically designed to assist data cleansing processes. Conversely, there are data discovery tools that are not data profiling tools.

Moreover, data profiling is closely associated with data quality but data discovery has far wider application than just data quality. For example, data discovery is important when implementing MDM (master data management) apart from its value in supporting data quality; it can be used to complement data modelling tools; it may be employed for business intelligence purposes; and it has a significant role to play in supporting data migrations, data archival and data governance, amongst others. For a detailed discussion of this topic see the Bloor Research Spotlight Paper on this subject that is being published to accompany this Market Update.

As a result of these two considerations: that data discovery isn't only provided by data profiling tools and that the utility of data discovery isn't limited to data quality environments, we believe that data discovery should be treated as a market in its own right.

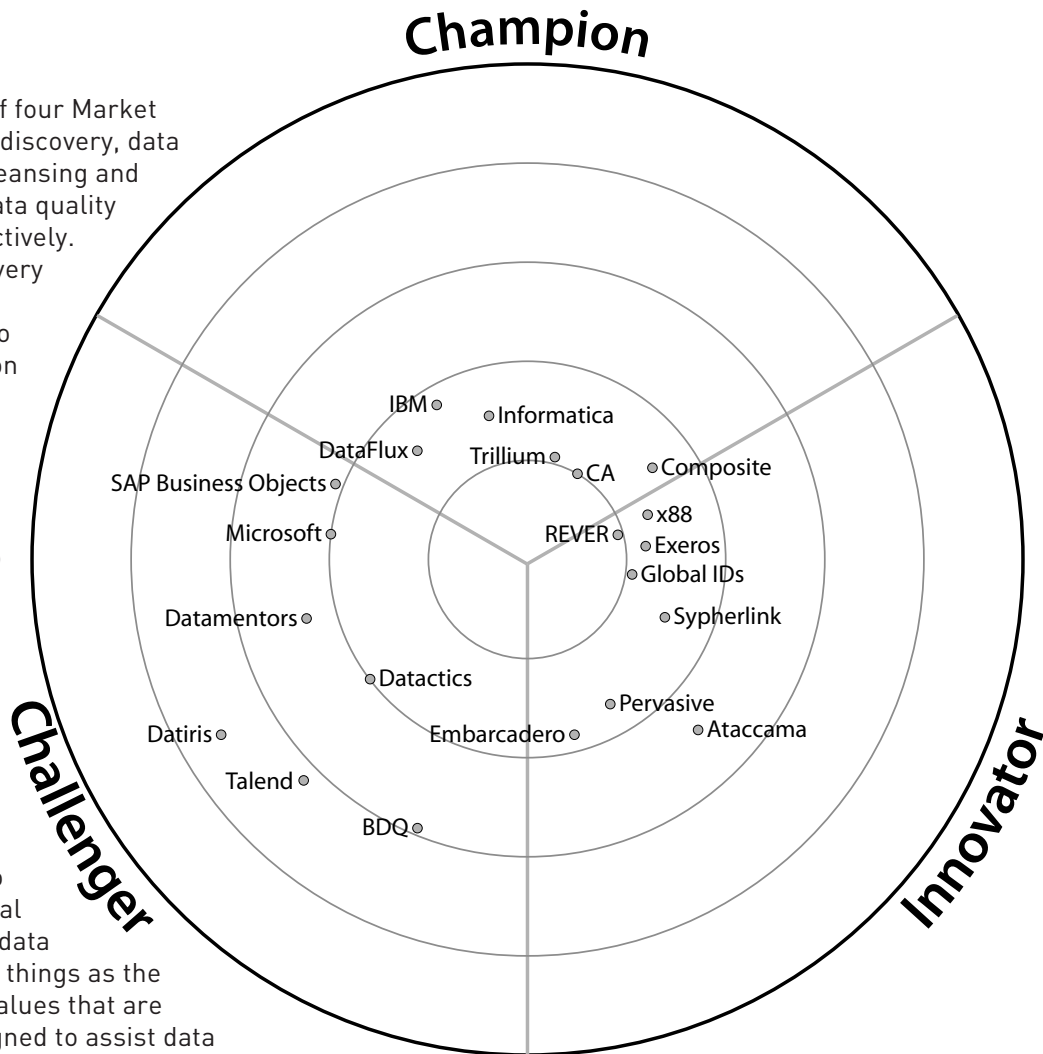


Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator segment if their innovation rating is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

Key market issues

There are, essentially, four different approaches to discovering data relationships. The oldest, and least efficient (at least if you working with more than one data source), is data modelling. Using tools of this type you can reverse engineer existing database schemas in order to discover the relationships defined within that schema. Unfortunately, many relationships are not defined within schemas so data modelling on its own is not sufficient to support data discovery. This highlights a major issue in that data discovery tools need to actually analyse the data itself and not just the metadata, because the metadata is inadequate. Moreover, since many relationships are defined within application software it will help if your chosen tool has an understanding of the relevant application environment (SAP, Oracle and so on) or programming language. To give an extreme

example, consider the COBOL redefine statement—this allows the same data field to be used for different purposes (for example, if the first digit is “1” then this data represents a customer, if it’s “2”, then...)—we know of one user that had 74 different re-definitions on a single field, which means that trying to determine relationships simply on the basis of the data and metadata will be hopelessly inadequate in these sorts of environments.

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.

These are not the only areas where many tools are deficient in functionality. Most, but not all, suppliers support profiling across multiple, heterogeneous data sources but the number of such sources that can be profiled simultaneously is often very limited. This will be a significant drawback if you are constructing an MDM hub from a large number of sources or if you are consolidating a number of different applications or databases. Another constraint is that relatively few vendors support sources other than those that you can connect to via ODBC, though most also support flat files. You can flatten XML using third party tools but you would prefer direct support. JDBC provides wider options but roughly one third of suppliers have yet to adopt this.

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 discovery 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.

Perhaps the biggest issue is automation. 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.

The third type of approach to data discovery is what we might call structured search. Here, the concept is that you have a federated platform that enables you to query multiple, heterogeneous data sources via virtual views. You build indexes (an automated process) against the tables (if a relational source) that you are interested in and then implement a search front-end. This sort of approach has the advantage that it can be used for both business intelligence and data discovery purposes. Similarly, it can be employed by both business analysts and data management personnel. This is a significant advantage since it is the business that has a better understanding of data relationships, and collaborative capabilities are important in this area. However, this approach will not provide some of the more advanced capabilities of data profiling such as outlier analysis or schema profiling.

In terms of collaboration more generally, 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 something that is rarely offered by data profiling vendors but should be standard for products in the structured search market that makes use of federated views. Partnering between data profiling and data modelling is another way to provide visualisation capabilities.

Fourth, there is one vendor that uses an approach to data discovery based on a model-driven architecture (MDA). That is, it reverse (and forward) engineers databases using physical, logical and conceptual models. This is especially important when migrating between database environments. One of the things that you want to do in data migrations is to compare the data in the new environment with that of the old environment and there are a few such tools available on the market. However, these only compare relational databases whereas, when using an MDA-based approach, you can compare hierarchical or network databases with relational ones. This manner of working can also understand COBOL redefinitions, as discussed previously, as well as which applications use which data, which is important because what may be good data for one application may be bad for another and you need to know which. It is also useful for chargeback purposes. Note that this sort of methodology will support the sort of things you would expect (but won't always get) from a data profiling tool such as identification of primary/foreign key pairs, overlap analysis, column redundancy, business and transformation rule discovery and exception discovery, semantic awareness to support business glossaries, and so on.

Finally, bear in mind the need for reuse. If you are a data modeller and you want to see if your metadata matches the data then you may have no further concerns. However, if you are engaged on an MDM project then you would really like to use the same tool to support overlap analysis and the discovery of matching keys as you do to support the data quality initiatives that will no doubt accompany the MDM initiative. On an a priori basis, where data quality is important for reuse purposes, then data profiling is likely to be most suitable. However, bear in mind that the approaches detailed here have reuse capabilities in other areas; so, for example, if you are starting your MDM implementation by deploying a registry (with a hub to come later) then you will probably need a federated engine that will underpin that solution that could also be used for structured search-based discovery.

Vendor landscape

There are more than 30 vendors involved in the data discovery space, marketing one or more of the technologies discussed. Of these, more than 20 responded to our requests for information while Sybase, Oracle and Human Inference declined to be involved, the first of these because it has yet to develop its plans for data discovery, Oracle 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 data discovery or profiling (though it does offer the latter).

Of the 21 vendors that we are therefore reporting on, two (CA (ERWin) and Embarcadero (ER/Studio)) are in the data modelling space, though the former has recently announced a partnership with Exeros and it will be re-selling that company's X-Profiler as CA Data Profiler (Exeros' more advanced offerings will be available as an up-sell from Exeros). In the structured search space Composite Software (Composite Discovery) is the only supplier that we know of that focuses on data discovery as well as business intelligence though, interestingly, one of the data profiling vendors plans to move in the opposite direction (using its discovered relationships as the foundation for business intelligence). REVER is the only vendor employing a model-driven architecture that addresses this market.

All of the other vendors offer data profiling solutions. However, there are three camps: 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), Datisis, Exeros, Sypherlink and x88; in the second are Ataccama, Datactics, Datamentors and Trillium; and in the third group are 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 in addition to the CA/Exeros reseller agreement already mentioned. For example, both BDQ and Datactics are embedded in third party data quality platforms. The same is true

of Ataccama whose technology is OEM'd 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 discovery offers. At least partly for this reason most of the leading products in this 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 vendor to seriously consider is CA, thanks to its partnership with Exeros, though CA will be focused on data discovery to augment data modelling rather than for other purposes. Nevertheless, CA is a clear leader, along with Trillium, amongst the major vendors.

Of the smaller players in the data discovery space we would single out Exeros, Global IDs, Sypherlink and x88 (which is very new and therefore has the potential for greater things) amongst the data profiling vendors, as well as Ataccama, particularly when used in conjunction with iWay's Integration Server for real-time processing. Pervasive is also worth a mention because of its high performance Datarush engine. Of the non-data profiling suppliers REVER stands out (even in comparison to the profiling vendors) and we would also like to commend Composite Software for its approach.

*Philip Howard
Data Management
February 2009*



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