Bloor MarketUpdate

Self-service data preparation 2015

Market basics

Data preparation is the art, or science, of combining data from multiple sources and preparing it for analysis. Historically, data preparation has been around for a long time but it was limited to relational data and required specialist expertise. What is now emerging is a market segment that is aimed at data preparation that not only spans multiple sources and types of data, but also provides self-service capability for the user. In most cases the user will be a business analyst though there are products that focus more on data scientists. Relevant products may also be suitable for use by developers or ISVs (independent software vendors) wishing to embed analytics into applications.

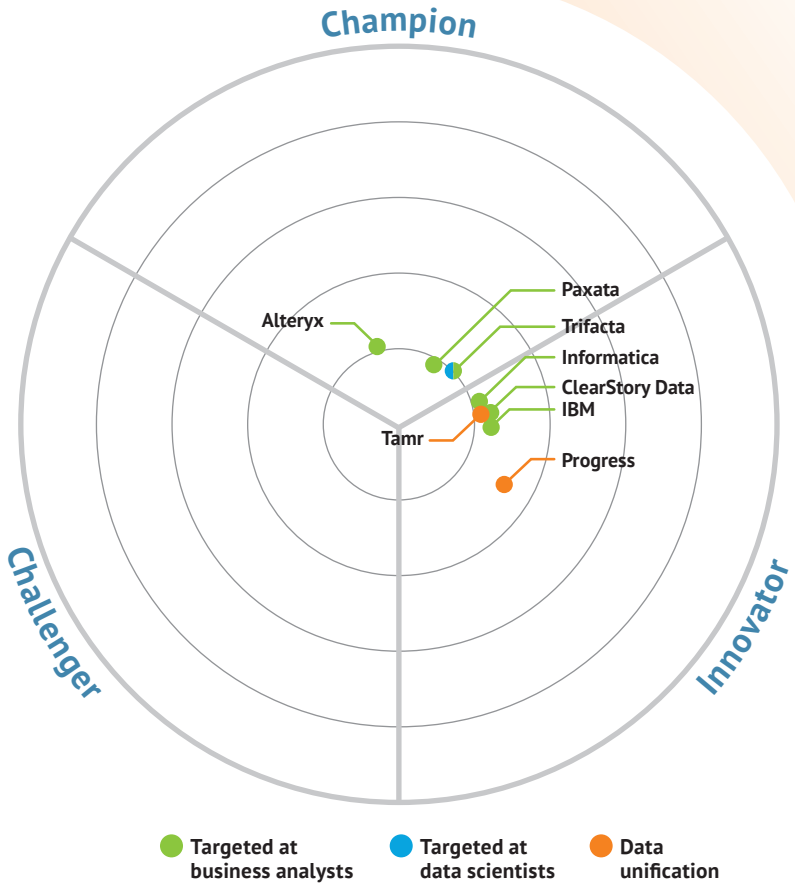
It should be noted that data preparation products are exactly that: they are about preparing the data and not analysing it and most offerings in this space integrate with third party business intelligence tools. This statement is not altogether true, however, as there are some vendors that provide a certain degree of business intelligence and analysis in their own right even though they may integrate with third party tools for things such as interactive visualisation.

As far as data preparation as a function is concerned, this means connecting to the relevant sources of data, joining data from these sources, de-duplicating it, transforming it, cleansing it, enriching it, filtering it, pivoting it, de-pivoting it and doing all of the other things that might be necessary prior to analysis. These are, of course, all of the sorts of things you would expect from data integration, profiling and quality tools. However, these traditional product types are IT tools, whereas data preparation tools of the type under discussion are targeted at end users. The question therefore arises as to how suppliers have achieved this?

The most common method used is to employ semantics and natural language processing to recognise common fields in disparate datasets so that the software can suggest to users how they might, for example, join these together and de-duplicate them. Combined with machine learning this will not only work out of the box but recommendations will improve over time. Alternatively, at least one vendor (Progress) is providing pre-built customisable templates for common joins (for example, Salesforce.com and Eloqua). Another vendor (Tamr) is providing pre-built vertical solutions for particular industry sectors. Semantics can also be used for other purposes: for example, to suggest when enrichment (such as demographic data) might usefully be added, or to detect invalid data. It is the appropriate use of these different types of technology: semantics, machine learning and artificial intelligence, or their equivalent, that has enabled the development of these self-service tools.

One final point is that data preparation platforms include auditing and monitoring capabilities. This is for various reasons. Firstly, it means that the use of data can be monitored by IT, both for compliance purposes and to give IT a better sense of what data users are accessing and how they are using it, which may enable IT to be more proactive and responsive to business needs.

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.



Secondly, if a business department decides that it wants to put a particular query (or technically, the preparation thereof) into formal production then all the logic of that preparation has been captured by the platform, so that the process of putting this into production will be much simpler and faster.

Market trends

The biggest trend is simply towards self-service data preparation in general. As yet there are a relatively small number of players though we expect more to come to market. Currently, the suppliers can be split into camps: major vendors and relative new boys. In the former category are IBM, Informatica and Progress while the latter category includes both start-ups and companies with a bit more longevity. As far as functionality is concerned there are three different foci at present: Trifacta is targeted at data scientists as well as business analysts, everyone else is more less aimed at business analysts and savvy business users at present; secondly, Progress and Tamr are perhaps better described as data unification platforms in that the focus is on unifying large numbers of different data sets more than on the actual blending of data, though this is really only a question of degree; and thirdly, ClearStory Data and Alteryx both provide some level or business intelligence or reporting as a part of their platforms, while the other suppliers do not. IBM's DataWorks is available either stand-alone or as part of Watson Analytics.

In so far as market trends are concerned the biggest one that we expect is convergence. We expect products to develop to the extent that they support both business analysts and data scientists and that they all encompass the unification of large numbers of datasets. We also expect more business intelligence and data warehousing vendors to introduce platforms such as those discussed here as well as companies that are active in data quality and governance. Indeed, we currently know of three other vendors that are palling to introduce products within this space.

Vendors

It is worth briefly outlining the distinctions between the different vendors covered by this Market Update. In alphabetical order:

Alteryx:

targeted at business analysts; part of the company's business intelligence offering (but partners with third party BI vendors for interactive visualisation) so that reports and so forth can be generated for executives.

ClearStory Data:

targeted at business analysts, includes storyboard generation for executives, otherwise you can export prepared data to third party BI and visualisation tools or you can import third party visualisations into a storyboard.

IBM DataWorks:

targeted at business analysts, also available as a part of Watson Analytics.

Informatica Rev:

targeted at business analysts.

Paxata:

targeted at business analysts.

Progress Easyl:

targeted at business analysts. More unification platform at present as it lacks some data quality functionality. The company will be embedding third party software for this purpose. Leverages DataDirect. Ships with out of the box templates. Cloud-based solution.

Tamr:

data unification platform targeted at business analysts but perhaps applicable to data scientists. Sees Trifacta (for example) as complementary. Offers a number of vertical solutions.

Trifacta:

targeted at both data scientists and business users. It has additional wrangling capabilities for the former.

To help to clarify matters we have colour coded the various vendors on our Bullseye Chart so that comparisons are between apples rather than between fruit types.

Conclusion

As noted, this is an emerging market and we expect more vendors to enter this market (indeed, we know of one other that is already planning that). A significant number of vendors: IBM, Informatica, Progress and Tamr have only launched their products within the last twelve months and in several cases much more recently than that. Indeed, Easyl, while available to the existing customer base (largely partners) is still not generally available. Needless to say, such early releases cannot usually be expected to be as rich in functionality as some of the more mature products on the market though vendors with a strong background in data quality and associated areas have a potential advantage in that they can reuse existing capabilities.

Bloor

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

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