Self-service data preparation & cataloguing

Market basics

In our 2015 Market Update on data preparation we concentrated solely on data preparation per se, which we defined as the art, or science, of combining data from multiple sources and preparing it for analysis. However, it is now clear that this is not the full extent of the capabilities that are required. Before you can blend data from different sources you need to be able to find and identify the data that is available to you and which is pertinent to the process in hand. This extension to data preparation would logically be called data discovery because, literally, that is what you are doing. Unfortunately, that term has already been taken so we are, instead, referring to this as data cataloguing.

Regardless of the type of product, there is a major emphasis on self-service. 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.

As far as data cataloguing is concerned, this means (automated) crawling through available data sources, collecting metadata about those sources and their contents, and then allowing users to search through the resulting catalogue to find relevant data. Data preparation represents the following step: 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.

Note that while some data preparation products are embedded within broader analytics offerings, data preparation is not about analytics per se.

A final point is that solutions 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. 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

There have been significant changes in this market over the last twelve months. The most notable is that whereas we included just eight vendors in last year's Market Update, this year we have considered more than forty products for inclusion and, in practice, have actually included half that number. In part this is because of the introduction of new data cataloguing providers but mostly because there has been a significant growth in established vendors moving into this market.

These established suppliers have approached data preparation from two different directions and in two different ways. Some are well-known players in the business intelligence and analytics space while others are providers of data integration and/or data quality solutions. Some of the former, and all of the latter, are marketing stand-alone tools that can be used independently of anything else that that company might offer. There are thus three classes of products within this space: BI-derived, data integration-derived and pure-play. ▶

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. Note that data preparation products, with/without analytics, have been directly compared with one another, but this is not the case with data cataloguing, with/without data preparation, which have been treated as distinct product sets.
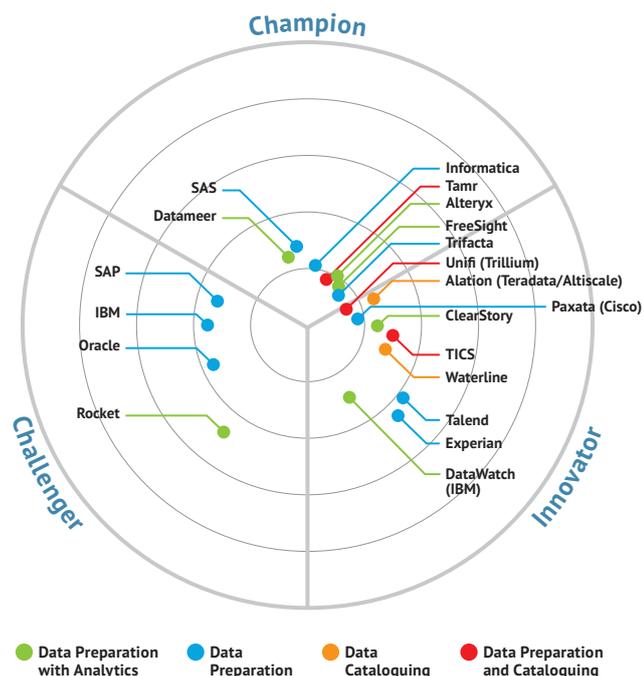
However, there is a certain amount of overlap as there are mainstream suppliers that have partnered with pure-play vendors and whose products they are reselling. In the longer term there must be a threat to the longevity of the pure players, simply because of the number of established competitors. Those that have already entered into significant partnerships should be well-placed to secure a profitable exit in due course.

Also within the context of market growth, we must comment that one company – Progress – has dropped out of the data preparation market and is no longer actively marketing its Easyl product, which appeared in last year's report.

We suggested last year that we expected convergence within the data preparation market. We still do. However, we now expect this within a broader context. Previously, there were two things that concerned us: broadening the reach of relevant products from data scientists to business analysts and vice versa, and extending integration capabilities. In the first case, this has happened and is continuing to happen. In the second case, we found that some products had relatively few connectivity options and we expected a broadening of that capability. Little has changed in this regard though a number of new entrants into the market have significant connectivity capabilities. In addition to these two areas of convergence we also expect to see data preparation and data cataloguing products subsuming each other's capabilities. In general, these are currently distinct products areas though there are a couple of exceptions.

There is one area of concern that is being addressed by very few vendors. Data scientists (and business analysts) developing analytic algorithms will often be in the same boat as conventional application developers in the sense that they are not authorised to see personally identifiable details within the data they are preparing. On the other hand, they do need realistic data to work with. This means that relevant data will need to be masked. However, as yet, relatively few suppliers in this market have recognised this need and they have neither developed relevant capabilities within their own products nor partnered with relevant third party data masking vendors. This is therefore not so much a market trend for today but one that needs to be initiated for tomorrow.

Finally, a word of warning: "data preparation" has become something of a bandwagon and a number of vendors are claiming to offer it even though they offer a paucity of required features. Our running count of suppliers claiming data preparation that don't actually fit the requirements of data preparation, is currently over twenty. All but one of these companies is a business intelligence vendor.

Vendors

In this section we briefly describe and discuss the various vendors and products covered in this Market Update, within a variety of categories. We think these categories are important to understand because they will impact on buying decisions.

Note that some products only run on Hadoop while others are only available as cloud-based services. While it is good to have these as options, they are limiting when there are no other choices provided. Similar considerations apply to products that are targeted specifically at business users and which do not provide facilities that a data scientist might want to use.

In the following, vendors named in brackets resell the products listed, in some cases, with extended functionality.

Data preparation and cataloguing – pure plays

Tamr

Tamr's initial focus was on what it described as a "data curation" platform. That is to say, it was and is especially focused on unifying and preparing data from a wide range of sources for downstream analytics or operational purposes. As this unification has been a relative weakness in some of the pure-play data preparation platforms this has led to Tamr being regarded as a potential partner by these other vendors. However, one would expect such suppliers to add further connectivity options over time so this would decrease Tamr's competitive advantage for analytic (not operational) use cases, had this remained Tamr's only focus. But, during the course of 2015 Tamr moved into the data cataloguing space, again providing capabilities that are not offered by the likes of Paxata and Trifacta. In the longer term, with other data cataloguing companies entering the market along with data integration vendors we expect Tamr to move to compete more directly with the pure-play data preparation vendors.

TICS Explora

Explora from TICS Services is part of the Uneartha Framework, which is a requirements-driven approach to data governance, data integration and, in the case of Explora, data preparation and cataloguing. The company has partnered, and integrated its product, with a number of third parties, include TIBCO Spotfire, Qlik, Tableau and OpenText (Actuate). The product is metadata-driven (necessary when it comes to discovering unknown sources of data) but does not directly leverage semantics. Nor does it have a recommendation

engine. On the other hand, you can mask data from within the TICS environment. Excel-like capabilities are provided for things like concatenation and column splitting, while SQL is available for more experienced users.

Unifi (Trillium)

Unifi is the only company that we are aware of that has set out from the beginning to deliver both data inventory and data preparation capabilities and which includes exactly the sorts of capabilities you would expect in both areas. It is built natively on Hadoop and dynamically leverages either or both of MapReduce Hive and or Spark depending on the nature of the transform job. It is targeted at both business analysts and data scientists with the ability to call various statistical algorithms from the product's Expression Builder. While you could argue about the individual features of the product there is no question that it has the broadest range of capabilities in the market. Unifi also provides overall IT management of the platform that ensures the IT organisation maintains control over, and user access to, data sources and Hadoop resources. No doubt, this is why Trillium is now re-selling Unifi. However, in this case, Trillium is not just reselling the product (as Trillium Prepare) but it also has an integrated solution called Trillium Refine with which you can deploy Trillium data quality processes (standardisation, cleansing, matching and so on) into your Unifi workflows.

Data cataloguing – pure plays

Alation (Teradata, Altiscale)

Alation describes its eponymous product as a data accessibility solution, which we are referring to here as data cataloguing. The emphasis in the software is on the discovery of data resources and the capture, and subsequent management, of the metadata relevant to that data. In other words, this is all about knowing what resources you have access to, which must be the starting point for data preparation. Features such as full text search are provided and the crowd-sourced documentation is praised by users. A further user comment was that the ability to assign and discover domain experts by table and by column is *"immensely useful"*. The product is being resold by Teradata with the two companies having a number of joint customers even before the partnership was negotiated. It is also noteworthy that Teradata is dropping Teradata Loom – its previous product in this space – in favour of Alation. Alation is also resold by Altiscale, which is a Hadoop as a service provider.

Waterline

Waterline Data is another vendor providing the ability to find, catalogue, provision and govern data resources prior to data preparation. Specifically, it provides a collaborative environment for business users to explore and find the best data, within a data lake, to accelerate data preparation. Waterline Data's technology automatically creates a catalogue of the data and presents that in an on-line marketplace-like interface: so that anyone that requires access can find it using the same approach they do today when shopping online. Like e-commerce marketplaces, additional benefits can be derived from this approach. For example, the catalogue can be augmented with tribal knowledge (crowd-sourcing) about data value and use scenarios. Further, with a marketplace approach, data governance policies can be applied to protect sensitive data, manage access, and increase trust in data quality and validity. Waterline partners with Trifacta and metadata discovered during wrangling with Trifacta can be used to augment Waterline's catalogue.

Data preparation – pure plays

Paxata (Cisco)

Paxata is built on top of Hadoop and Apache Spark. Above this is an application web services layer where you establish projects, enrich the data with external data, look at the distribution of different types of data, match and cleanse data, and so on. There is an in-built automation engine enabled by machine learning, semantic indexing, statistical pattern recognition and text analytics techniques. This handles data in a model-free environment and operates over a large variety and volume of both structured and unstructured data with real-time capabilities enabled by a vector query processor. Paxata includes semantic capabilities that will recognise likely equivalences between data in different data sources and it will make suggestions as to things such as join keys. At the front-end there is bi-directional integration so that it is possible to move between data preparation projects to a BI product (e.g. Tableau) and back without leaving your BI environment. Paxata and Cisco developed a joint solution as a part of Cisco's analytic strategy for data centres, data virtualization and the Internet of Things.

Trifacta

Trifacta offers both Trifacta Wrangler Enterprise and Trifacta Wrangler, with the former being Hadoop-based and the latter being a desktop product that does not require Hadoop. Wrangler Enterprise can be deployed in the Cloud or on-premises and supports multiple execution frameworks include Apache Spark,

MapReduce and single-node execution. In addition, the company has just launched (March 2016) the Photon Compute Framework which incorporates high-performance in-memory capabilities (compatible with Apache Arrow) directly into the Trifacta interface. From a functional perspective, it is probably beyond question that Trifacta is the leading data preparation product for data scientists, with a specialised language – Wrangle – to support these users, although the company states that the majority of its users are business and data analysts rather than data scientists. The company offers different interface options for users including Visual Cards, a menu-driven Transform Builder and Wrangle itself. While to our eyes even the Visual Cards are not that intuitive, users seem to disagree. For example, one commented to us that *"I love that you don't have to be a coding expert or tech genius to understand the product"*. Usability is very much in the eye of the beholder. From a functionality standpoint Trifacta stands comparison with any other product on the market. Notable partnerships – apart from those with analytics vendors – are with Waterline and with Cloudera Navigator, where the former provides data cataloguing and the latter offers data governance for Hadoop. In both cases Trifacta can exchange metadata with the partner product.

Data preparation – hybrid BI/stand-alone

Products in this category may include business intelligence and/or analytics capabilities. However, where BI capabilities are included they are often used only for preparatory analysis and they are frequently integrated with various front-end tools such as Qlik and Tableau.

Alteryx

Alteryx started life with basic data preparation and spatial analytics capabilities, and has since added more advanced data preparation capabilities, as well as predictive analytics based on R. It would now be true to say that Alteryx is a data preparation vendor with advanced analytics capabilities. Alteryx is deployed on a Windows desktop OS, and is typically used by data analysts in a line of business departments such as Marketing, Operations, or Finance. From a functionality standpoint, the biggest strength of Alteryx is its workflow capabilities where you build the processes that lead through from initial connectivity (to data warehouses, cloud applications, spreadsheets, and many other sources), to data preparation, data blending, and advanced analytics. This is a highly visual process, as opposed to a number of other vendors that either do not have workflow capabilities or have them but via a less intuitive interface. Users we have contacted have universally praised this feature and the way that it saves time. It is often the case that results are outputted from Alteryx to tools such as Tableau, Microsoft or Qlik for data discovery and visualisation.

ClearStory Data

ClearStory Data is an end-to-end cloud-based analytics solution providing data access, machine-based data preparation via "Smart Data Inference", data blending via automated "Data Harmonization", visual data analysis, data sharing, and collaboration on insights. The solution's strengths include its data profiling and inferencing, and automatic data blending (or harmonisation) capabilities, all leveraging a data processing framework in Apache Spark. These capabilities are critical especially for use cases where non-technical users must bring together multiple data sets to arrive at insights in a timely manner. Machine-driven "Data Inference" (data preparation) captures the semantics, cardinality, lineage, and the associated metadata from each data source that is accessed in an analysis to create a logical in-memory model of the data. Data Inference is distinct from data cleansing in that Inference is designed for reaching data semantics to speed data modeling whereas cleansing is about correcting dirty data. Business users consume analysis via ClearStory's interactive and collaborative StoryBoards. These StoryBoards are living storylines that update as the underlying data refreshes, while maintaining context on the insights (state, lineage, tasks performed, the user's exploration path and so on) for every insight can be tracked as data changes.

Datameer

Datameer started life by providing a spreadsheet interface to Hadoop for business intelligence and analytic purposes, and now provides an analytics solution that includes full-blown data preparation. Datameer leverages Hadoop for both storage and computational purposes, and is available either in the Cloud or on-premises. It retains its spreadsheet-based approach to provide a familiar interface for business users. Unlike most other vendors in this space, the company has addressed security issues with built-in data masking and through leverage of Hadoop's encryption-at-rest capability. From the perspective of data preparation specifically, there are a number of notable features – such as the ability to calculate the distance between locations automatically – as well as pre-built functions for handling complex data types like URLs, JSON, and XML. The product makes significant use of semantics but, unlike some other vendors, Datameer does not make suggestions as to how you might join different files. Because of the significant scale of data that Datameer works with, data preparation and analytics design is built on

the basis of a "smart" sample of data, and as a result, recommendations might lead to anomalous results

Datawatch Monarch (IBM)

Historically, the strength of Datawatch has been the ability to process not just conventional data but also unstructured data: to extract (numerical) data from generated reports and documents. In this sense Datawatch was a "big data" company not just years but decades before the term was invented. In 2013 Datawatch acquired Panopticon and its eponymous product (now Datawatch Designer), a visualisation vendor specialising in the monitoring and analysis of real-time data. As an alternative Datawatch has a partnership with Tableau. Datawatch Monarch is available as a Windows-based, on-premises solution and it connects to a wide variety of sources (one of the product's strengths), including Hadoop, as well other NoSQL environments. Data governance, quality and profiling are all included within the product and Datawatch is one of the very few vendors in this space to provide data masking. Machine learning capabilities and a recommendation engine are currently in development. The company has just signed an agreement for IBM to resell Datawatch Monarch in conjunction with IBM Watson Analytics and IBM Cognos Analytics.

FreeSight

The following is an extended description of FreeSight, commissioned by FreeSight Software. In all other respects this paper is identical to the generic Market Update published by Bloor Research.

FreeSight is currently in version 3.1. It is a 64-bit Windows-based product and it has an interface that looks and feels like – but is considerably easier to use than – Microsoft Excel. For example, you have to be a pretty experienced Excel user to deploy and manipulate pivot tables whereas FreeSight's equivalent requires far less expertise, while retaining the sophisticated capabilities that expert users might expect. The product is available on a 30-day free trial and subscriptions (which include maintenance and upgrades) are just USD $295 per annum per computer. There are effectively three different markets for FreeSight:

1. As a stand-alone self-service data preparation platform, potentially linked to a third party front-end visualisation tool.

2. As a data collection, preparation and reporting platform, that goes all the way from data capture through data integration and cleansing through to reporting, with some (although not extensive) data visualisation capability.

3. As a platform for capturing and automating what would otherwise be repetitive and error-prone Excel processes.

Specifically, FreeSight is two things: an end-user tool for data preparation and analytics, and, for more advanced users, a development tool for automating reporting and other business processes. The company's users claim significant time savings compared to Excel and that the product is "much more visually appealing and easier to understand from a visual perspective, and is much easier to explain to a non-user or Excel user". Specifically, users work in FreeSight via a visual canvas where all data manipulations are performed, resulting in a process workflow that is live (each node in the workflow provides direct access to underlying data and operations) as opposed to just a picture. This supports both data governance and auditing, as well as the ability to restore to a former state. It is a significant differentiation compared to some other data preparation tools in that with FreeSight you can drill down from nodes in the process flow (other products are often static), and FreeSight also supports the ability to reverse any operation at any time.

From a data preparation perspective FreeSight has extensive semantic capabilities though no machine learning. On the other hand, the company has a number of patents to its name with respect to automated joining and cleansing (the software automatically analyses source data for relationships), as well as strong inferencing and profiling capabilities, all based on the product's semantic strengths.

With respect to Excel, FreeSight supports formulae in much the same way but has an extended catalogue of capabilities with some 80 additional capabilities for things like date and time, text manipulation and so on. The product has auto-charting capabilities that are much easier to use than pivot tables and which makes this sort of capability accessible to business analysts who are not Excel experts. Another notable feature that is lacking in Excel but present in FreeSight is the ability to analyse across different versions of the same spreadsheet (for example, where you have each monthly set of data stored under a different tab).

IBM DataWorks

IBM DataWorks is a fully managed offering for self-service data preparation that leverages Apache Spark and runs within IBM Cloud Data Services. It integrates with Watson Analytics as well as with IBM Cloudant, IBM dashDB and DB2 on Cloud, and it connects to a wide variety of other sources. It is targeted primarily at citizen analysts that need to get data ready for analytics, whether that is for reporting, visualisation, modelling or other analytical tasks. IBM is also reselling Datawatch Monarch as an on-premises solution.

Oracle Big Data Preparation Cloud Service

As its name implies this is a cloud-based service. It is available in the Oracle Cloud and is based on Hadoop and Apache Spark. The "big" in the title is redundant since we would regard any so-called data preparation product that was not capable of handling unstructured and semi-structured data as not worthy of the name. The product includes both statistically-based machine learning capabilities and semantics (natural language processing) that support recommendations. The company is claiming that this is a differentiator for it but we beg to differ, as there are a number of other products that have both these capabilities although, to be fair, there are also a number that don't. While we can understand why a small vendor might offer a cloud-only service we wonder what Oracle users that want an on-premises capability will do.

Rocket Discover

Rocket Discover is a self-service data preparation, visualisation, and discovery product. While Discover can easily connect to and blend data from multiple sources – ranging from mainframes to databases to social feeds, initial go-to-market focus is aimed at IBM data sources and those offered by Rocket itself. For example, Discover explicitly supports the U2 (UniVerse and Unidata) multi-value databases as well as IBM Cognos TM1 and Cognos BI. With a focus on TM1 that leverages a RESTful API, Discover uniquely supports critical TM1 features such as aliases, attributes and the ability to write back to TM1. While Discover will certainly be appealing to existing Rocket users, the company also expects significant interest outside of that market.

SAP Agile Data Preparation

The SAP Agile Data Preparation application runs on the SAP HANA Platform, whether on-premises or in the cloud. While the features and functionality of SAP Agile Data Preparation are extensive this is the only software in this space that requires that you license additional technology to support your tool (SAP refers to it as an application) of choice. Note that although this is the SAP HANA Platform this does not necessarily require the use of SAP HANA because that database is not mandated by the platform, despite its name! This is sufficiently confusing that we expect that SAP will only be marketing the Agile Data Preparation product to existing SAP customers.

SAS Data Loader for Hadoop

As its name implies this is currently a Hadoop specific product that runs with either MapReduce or Spark and HiveQL or Impala. Data is loaded into Hadoop from relational databases, text files or SAS datasets, with full data profiling and data quality processes, as well

as SAS code functions, running in parallel within the Hadoop environment. For analytic and visualisation purposes you have to export prepared data back to the source database or in-memory into the SAS LASR Server. In all of these cases SAS plans to extend these options at some point in the future. Currently, while the facilities provided are sophisticated there are also limitations on the product. The company recognises this and has substantial plans for the future, including the introduction of a recommendation engine, cognitive machine learning capabilities and a more graphical user interface. Extended governance and data masking capabilities are also planned. Since the most recent release it is now possible to build repeatable workflows but a drag and drop interface would be more intuitive. For existing SAS users, a notable feature is that the product generates SAS code under the covers, which you can customise.

Data preparation – hybrid data quality/data integration/stand-alone

Note that Trillium's offering, which would otherwise appear here, is discussed under the UNIFi heading. It would also be reasonable to include IBM, Oracle, SAP and SAS under this heading. The big advantage that products in this category have is that the processes developed by business analysts can be operationalised in an automated fashion.

Experian Pandora

Experian Pandora is the only product in this update that has not been built or extended specifically for data preparation. Instead, Pandora has rather serendipitously evolved from its origins as a data profiling and quality tool by adding transformation capabilities that enable it to play in this space. In particular, the product has always had a reputation for ease of use and it is frequently employed by end users. In addition to its profiling capabilities – which have always been particularly strong in supporting heterogeneous environments – the product also includes the sort of workflow capabilities that you get from Alteryx, albeit that in the case of Pandora these are not graphical. Experian is now explicitly targeting this market and extending its platform to provide more features that are specific to data preparation. While the company has already embedded an engine that supports R it needs to add a number of other features to make it more closely resemble (or surpass) pure-play and other data preparation products.

Informatica Cloud – Rev

The initial offering of Informatica's collaborative data preparation platform, Rev is a cloud-based solution running on AWS in multi-tenant mode, leveraging a columnar data store. Using data flow wizards, IT can

create shared data connections to disparate business applications and databases and provide federated access to this data to business users. Business users access data through an Excel-like interface and further filter, blend, aggregate, formulate, cleanse and enrich data. Any processes applied by business users can be published down into the data integration interface as data flow maps so that these can be operationalised. At the front-end, Informatica has partnered with both Tableau and Salesforce Wave Analytics. A recommendation engine backed by both machine learning and semantics is included within the product. Currently available in a SaaS model, this overall architecture will start being available on-premises starting in May 2016 with extended big data preparation capabilities featuring an Enterprise Information Catalog enabling users to access, extract, prepare and store data to/from a Hadoop data lake.

Talend Data Preparation

One thing we especially like about Talend Data Preparation is the user interface which, in our opinion, is one of the best we have seen in terms of ease of use. That said, the product is relatively new and is not yet fully featured. For example, only outer joins are currently supported. However, Talend is an open source vendor and it has an aggressive schedule for building out the features of the product. In practice, you manage "recipes", which are sets of preparation tasks that can be shared, maintained, replayed and integrated into a data flow that can be managed through Talend Data Fabric for operationalisation. There is a recommendation engine with machine learning and semantics to support it. Currently the product is only available on-premises but Cloud is planned for later in 2016, as is support for data stewards (in addition to the currently supported field workers and business analysts), data masking, the leveraging of Apache Spark and numerous other features.

Data preparation – embedded BI

We had originally intended to include a variety of products from within this category. Indeed, we investigated a number of such solutions in detail before we concluded that this would not be appropriate. There are two reasons for this. Firstly, just about every business intelligence and analytics vendor in the world claims to have data preparation capabilities. It is easy, for example, to visualise data profiles and therefore to make data preparation claims. However, in our view, these claims are frequently spurious and a large proportion (not all) of analytics vendors do not actually deliver self-service data preparation capabilities or, at least, anything substantial. Exceptions to this rule would include MicroStrategy, RapidMiner, Platfora, Dell Statistica

and others such as Birst, Logi Analytics and Advizor Solutions. Secondly, we do not believe that most users will license any of the products in this category purely for their data preparation capabilities, but rather for the total set of capabilities that are offered, which will be focused on insight and analytics in the first instance. For example, we do not think that you would license the MicroStrategy Enterprise Analytics Platform just because you like its data preparation capabilities. We have therefore excluded products in this category from this update.

Conclusion

As we have discussed, we believe that this market will converge, both in terms of functionality (preparation, connectivity, cataloguing) and audience (business analyst and data scientist). After capabilities of the various products the major talking point is pure-play versus analytics versus data integration. The first of these has the advantage that it doesn't tie you to anything else, the second has the advantage that business users can stay (most of the time) in one environment, and the third has the advantage that you can easily operationalise the processes that have been developed. Thus this is not quite a best-of-breed versus platform argument because there is also the question of which type of platform?

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 different fruit types. Not surprisingly, the best scoring products within each category, are those that have been in the market the longest. Companies such as Alteryx, FreeSight, Tamr, Trifacta, and Paxata all score well and they are joined by Unifi as a highly ranked newcomer. We regard both Alation and Datawatch as ones to watch for the future, along with Experian and, especially, Talend. Honourable mentions should go to both Datameer and ClearStory Data. With the exception of Informatica and, possibly, SAS, we think that the mainstream vendors are somewhat behind the curve at present.

Philip Howard
Research Director
August 2016