Pure-play Data Integration

MarketUpdate

Market Basics

Data Integration tools first started to appear in the early 1990s. In other words the market is more than 25 years old. Needless to say, in terms of its basic capabilities the market is mature. However, that doesn't mean that it is static. Indeed, there is a distinct split between the vendors in this space. On the one hand there are vendors promoting a broad range of capabilities – typically including data quality, data governance and data cataloguing, amongst other things – as a platform, while on the other there is a second group of suppliers that have stuck more closely to their roots, and which we are here describing as *"pure-play"* data integration vendors. Trying to compare these different approaches in a comparative analysis such as this would be invidious: like comparing apples with fruit bowls. As a result, Bloor Research is publishing two separate reports on this topic, one focusing particularly on platforms (see ***www.Bloorresearch.com***) and this one, which concentrates on companies in the pure-play category.

Note that just because a vendor is currently characterised as pure-play doesn't mean that it will remain so, and it is likely that some of the some suppliers in this Market Update will transition over time. It also doesn't mean that the existing capabilities offered are limited to just data integration. For example, API and application integration, document processing, data virtualisation and data warehouse automation are just some of the extended capabilities offered by suppliers in this report.

Finally, it is worth considering why pure-play vendors continue to thrive. After all, a platform-based approach is attractive, but they are not always as well integrated as they might be, they will often be relatively expensive, they may lack features in particular areas that are important for a required use case, or you may prefer not to be dependent on a single provider. No doubt there are other potential reasons.

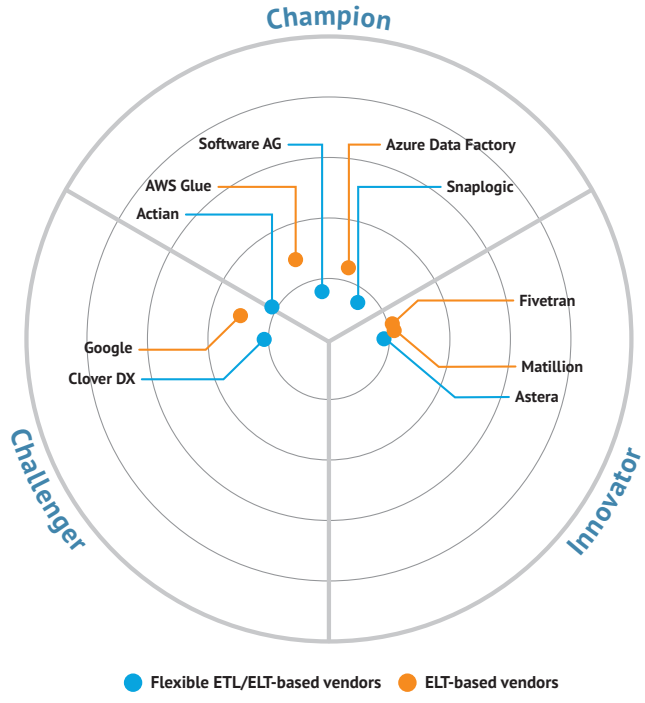

Figure 1: The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator or Challenger segments, depending on their innovation score. The exact position in each segment is calculated based on their combined innovation and overall score. It is important to note that colour coded products have been scored relative to other products with the same colour coding.

Market trends

Most of the trends in this market are generic – that is, they don't just apply to data integration tools, pure-play or otherwise - and we will come to these in a moment. However, there is one clear trend within the market itself which, in turn, reflects broader changes within the IT landscape. Historically, data integration tools were about moving data from one or more sources (usually databases, but sometimes Excel or other such environments) into a data warehouse or data marts. Today, an application environment such as Marketo or Salesforce is as likely to be a target as a data warehouse while pre-existing on-premises applications are as frequent as a source. Note that you are still moving data – so data integration is the correct terminology – but not necessarily between databases. Further, we now have data lakes, so integration based on unstructured data is an increasing requirement. So, data integration is no longer simply about structured data.

The underlying driver for the above is the move to the cloud and the other major element to this trend is in the growing adoption of cloud-based data warehouses. This often, though not always, involves migration from an existing on-premises data warehouse to a new one, from a different vendor, that is cloud-based. Supporting this migration is a major focus for data integration vendors of all stripes.

Cloud-native

This is the first of the generic trends, in the sense that all software providers are having to make this transition. There are multiple definitions of *"cloud-native"*. For example, the Cloud Native Computing Foundation essentially defines it as being based on the user of containers and their orchestration. Others take a broader view. Our preferred definition is that software is cloud-native if it *"exploits the technological and economic benefits of cloud-based computing that would not generally be available in non-cloud environments"*. Thus additional capabilities such as serverless computing, elastic (preferably auto-) scaling of resources (storage, compute and others) as well as consumption-based pricing. All the vendors in this report are moving in this direction, if they have not done so already. The trend towards providing managed services is also evident in this market.

Automation

Automation has always been a factor. Indeed, that's what computers are for. But within the context of data integration it has been a perennial concern: it makes life easier, supports self-service, reduces costs, and improves efficiency. Today, there is an increasing emphasis on AI and machine learning and it is certainly true that these technologies can introduce automation in a variety of ways. However, this does not mean that machine learning is necessary to support automation. As an example, machine learning can be used to make recommendations of various types. But that's not the only technology that can be used for that purpose. What we are seeing as a trend in this market is increased automation. It is often, but not exclusively, supported by machine learning and AI. For those implementing machine learning some, but not all, are providing facilities that explain results. We would like to see other vendors pursuing this course.

Vendors

The one major distinction that exists across pure-play data integration providers is their approach to integration. Traditionally, this has been via extract, transform and load (ETL) but with companies moving to cloud-based analytic environments there are an increasing number of vendors that only offer bulk load (suitable for some migrations) or ELT followed by change data capture (CDC). These approaches lack some of the capabilities of traditional data integration tools (for example, they don't support the B2B exchange of data and wouldn't support a TEL approach that would be suitable for working with blockchain), which typically offer more flexible environments with ETL, ELT and TEL capabilities, and combinations thereof, as well as CDC and streamed data. In our Bullseye diagram we have highlighted this distinction by colour-coding ELT vendors differently from those suppliers offering broader capabilities.

As far as the vendors and products considered in this Market Update are concerned, we have focused exclusively on those that are not offering a broad platform. The one product that we would have included, but have not, because it did not respond to our requests for information, is Boomi. Both Amazon and Microsoft provided us with relevant documentation but without giving us a formal briefing, while in the case of Google we were forced to rely on publicly available information. In all other cases we have had significant interactions with the various suppliers.

As a final note, it is worth commenting on hand coding. Most vendors still put this at the top of their list of competitors. We continue to be surprised by this as you get no reuse, no self-service, no automation, and no integration with other necessary technologies. Any upfront savings are false economies compared to the extra costs associated with rework, administration and other expenses that you don't get in a tool-based solution. Hopefully, the increased availability of managed services and consumption-based pricing will see off those users that still think that hand coding is a good idea.

Metrics

We have identified eight core capabilities to evaluate the products included in this report. For any given product we have considered how well, and to what extent, each of these capabilities is supported.

- **Structured connectivity.** Native connectors to relevant sources and targets are to be preferred, not least because they provide improved performance. In the old days, it used to be – more or less – sufficient to support ODBC/JDBC as generic connectivity options but with the growth in data volumes and increased demand for low latency, these are becoming less and less viable though they may still be useful in some instances. Note that the number of connectors claimed by vendors can be deceiving. For example, it is common to have six or eight or ten different connectors for each source, depending on the function of the connector. And, of course, different vendors count these connectors in different ways, so you need to be clear that you are comparing apples with apples. An additional question is how you integrate with sources and targets not supported by your supplier. A software developer's kit (SDK) that allows you to develop your own connectors will be useful as will be a vendor-provided portal where users can share the connectors that they have developed.
- **Semi/unstructured connectivity.** Connectivity is such an important capability for data integration tools that we have separated this into two metrics. While many of the same observations apply to semi and unstructured data as to structured data the ability to parse pdf and Word documents, to enable B2B exchange through support for standards such as HL7 or EDIFACT, or to support integration with edge devices and sensors goes a significant step beyond merely moving data into a data warehouse or lake. Built-in support for formats such as Lidar and video will be important in some Internet of Things (IoT) environments.
- **Flexibility.** This is a measure of the breadth of capability offered with respect to different approaches to moving data. In other words, does the vendor support both ELT and ETL? Does it provide bulk loading, change data capture, and support for streaming environments such as Kafka and Flink. If it does support both ETL and ELT does it offer push-down optimisation? If it supports streaming does that include both continuous and event-driven streaming, or only one of these?
- **Transformations.** How extensive are the transformation capabilities provided? Are there pre-built transforms provided and, if so, how many and for what purposes? In the case of ELT environments do these run in-database (which will improve performance)?
- **Workflows.** Are pre-built workflows or workflow templates provided? If so, how many and for what purposes? Can you embed one workflow inside another? What facilities are provided to enable the reuse of workflows? Is there version control? Are there suitable facilities for reporting errors and/or any workflow failures? Are there testing facilities provided?
- **Architecture.** This relates to the way in which the product is deployed, in particular whether it is cloud-native (as discussed previously) but also to what extent it supports traditional virtues such as performance (parallelism is important here), scalability, resilience, security, job scheduling, and so on. Idempotency is a relevant factor. In addition, our architecture metric includes consideration of extended capabilities that vendors may offer: data warehouse automation for example, or data virtualisation.
- **Ease of use.** There is an increasing trend towards the provision of self-service capabilities, which implies the use of no-code visual (drag and drop) development environments. Some products have separate interfaces for citizen developers and IT developers, but we would prefer to see a single user interface because this will enable greater collaboration. Other features enabling ease of ease include out of the box accelerators and similar constructs, as well as the general look and feel of the product.
- **Automation.** Self-evidently important, the automated capabilities present in a data integration tool can be a significant differentiator, and in many ways is a matter of *"the more, the better"*. The most highly automated solutions will feature a variety of embedded automation and machine learning throughout their built-in data processes and may even provide extensible automation capabilities as well. It is notable that some vendors are significantly in advance of others when it comes to the implementation of automation through machine learning. However, even without the use of machine learning, features such as workflow and process automation will be welcome.

Conclusion

Opting for a pure-play data integration solution may be appropriate for a variety of use cases but the available offerings essentially break down into two groups: those focused on cloud (and hybrid) environments, with solutions based around ELT processes; and more generic offerings. The former aim to out-compete platform-based solutions by doing a limited number of things extremely well, while the latter simply aim to be better, or broader in scope. In both cases, pure-play vendors will typically claim significant total cost of ownership benefits when compared to the big boy platform vendors.

About the authors

PHILIP HOWARD
Research Director:
Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

DANIEL HOWARD
Senior Analyst:
Information Management and DevOps

Daniel started in the IT industry relatively recently, in only 2014. Following the completion of his Masters in Mathematics at the University of Bath, he started working as a developer and tester at IPL (now part of Civica Group). His work there included all manner of software and web development and testing, usually in an Agile environment and usually to a high standard, including a stint working at an 'innovation lab' at Nationwide.

In the summer of 2016, Daniel's father, Philip Howard, approached him with a piece of work that he thought would be enriched by the development and testing experience that Daniel could bring to the table. Shortly afterward, Daniel left IPL to work for Bloor Research as a researcher and the rest (so far, at least) is history.

Daniel primarily (although by no means exclusively) works alongside his father, providing technical expertise, insight and the 'on-the-ground' perspective of a (former) developer, in the form of both verbal explanation and written articles. His area of research is principally DevOps, where his previous experience can be put to the most use, but he is increasingly branching into related areas.

Outside of work, Daniel enjoys latin and ballroom dancing, skiing, cooking and playing the guitar.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

tel: **+44 (0)1494 291 992**
web: **www.Bloorresearch.com**
email: **info@Bloorresearch.com**