(Cloud) Data Management (Platforms)

MarketUpdate

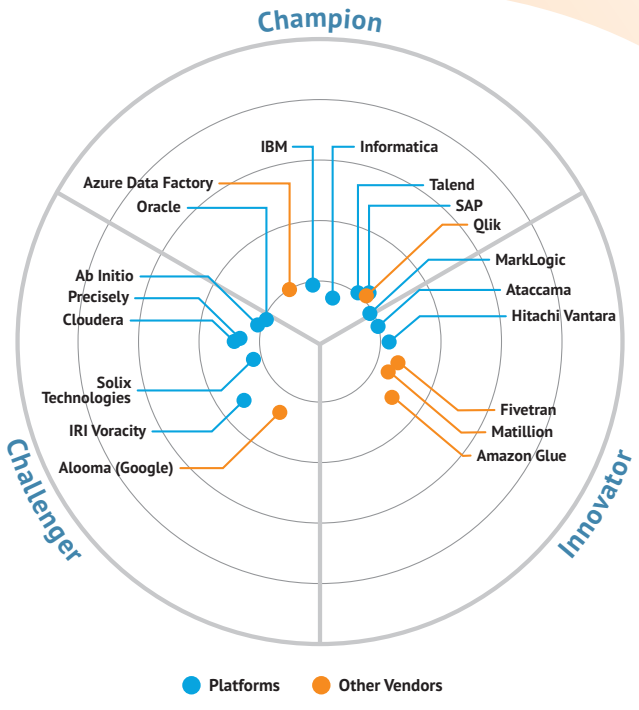
Market Basics

We should explain why we have put (cloud) and (platforms) in parentheses in the title of this Market Update. Firstly, we do not see that it is necessary to have a cloud-based data management solution to support cloud-based data warehouses and lakes or, indeed, other end points. We agree that it would be nice. We agree that it has advantages. But it isn't necessary. This is why we have put *"cloud"* in our title in brackets. Similarly, why have we also put *"platform"* in brackets? The reason is because there are a number of relatively new cloud-based data integration offerings that would not, in our opinion, qualify as fully-fledged – in some cases they are barely out of the nest – platforms. Nevertheless it is clear that significant numbers of users are adopting these as alternatives to the platforms we have set out to evaluate, and that therefore we should include leading products in this category as a part of this report.

So, what do we mean by a *"data management platform"*, in this context? There are five requirements that we believe to be mandatory. These are, first, some sort of data integration capability to get data into your target environment. Second, you need the ability to ensure the quality of your data, so that it is fit for purpose; third, you need to be able to apply data governance policies to your data and ensure they are implemented; fourth, you need to be able to identify and handle (mask) sensitive data to comply with appropriate regulations; and finally, you need to be able to support metadata management, typically by leveraging a data catalogue. *"Platforms"* in this Market Update all provide data integration capabilities and at least three of the four other requirements.

We should also point out that there are a number of complementary capabilities that are not fundamental to a data management platform, but which would be nice to have. These include support for master and reference data management, data preparation, information lifecycle management along with data retention and archival, data warehouse automation, and data virtualisation. Some products extend more deeply into analytics or privacy (test data management). Support for these sorts of capabilities are not reflected in our assessment of the various offerings surveyed for this Market Update.

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.



SAS has asked not to be included at this time.

Market trends

This Market Update makes the assumption that everyone is moving to cloud-based analytics platforms. Evidence suggests this to be true. However, there is the supposition, fostered by vendors that should know better, that moving from on-premises to cloud-based environments is a panacea whereas, in practice, it is just another deployment option. That moving to cloud is what you need to enable digital transformation. We do not agree. It should certainly provide more flexibility and agility, and it ought to provide that more cost effectively. However, if you are looking to move to a more data-driven environment then it is the data that is important not where it is stored. There are a few vendors in this report that have recognised this and are focusing on data integrity or data assurance but all too many are still concentrating on their technology as opposed to what is important for business users.

More generally, there are a number of trends in this space that have already been mentioned and there are still others that need further discussion.

Cloud-native

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, and [possibly multi-]compute) as well as consumption-based pricing. Many of the vendors in this report are therefore cloud-native from a technical perspective though not necessarily an economic one. Those that are not now invariably plan to be.

However, there is a caveat to this: data quality processes such as matching and de-duplication require that they be run within the user's computing environment rather than in that of a vendor. This is because data matching, by its very nature involves sensitive data and it would be non-compliant (with GDPR, CCPA and so on) to extract this information into a third-party cloud environment.

Automation

Automation has always been a factor. Indeed, that's what computers are for. But within the context of data management 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. We are also seeing companies leveraging machine learning putting a significant emphasis on the explainability of the algorithms being used. While there is not enough evidence to suggest that this is a trend it is certainly something that we look to see become one.

Integration

Another major trend is integration. There are actually three parts to this: data ingestion, data access/exposure, and interoperability. The first two of these primarily refer to how many data sources and third-party products (respectively) the platform supports (and also how they support them). While historically the former used to be about database support but there is now much more emphasis on supporting application environments, including those that leverage semi-structured and unstructured data (for example, Slack or Twitter). For IoT environments there is a greater need to support various types of sensors, LiDAR, video and other more esoteric data sources. As noted previously, targets have also broadened out, and loading into, say, Salesforce is increasingly commonplace.

Interoperability is perhaps more interesting, referring to how well the platform integrates products and services between themselves. In the best cases, this will reach further than simple interface connectivity and into active synergy, wherein the capabilities of two or more products are used in concert to multiply each other's strengths. In essence, interoperability is what separates a platform from a collection of disparate products. A metadata management layer shared by all products within a platform is arguably the best way of achieving this.

How to integrate

Related to this point is the actual process of data 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 migrations) and/or ELT followed by change data capture (CDC). These approaches lack some of the capabilities of traditional data integration tools, which typically offer more flexible environments with both ETL and ELT capabilities, and combinations thereof, as well as CDC and streamed data support. A third alternative is to use a model-driven approach to data integration whereby you map your sources to a relevant model and then map from that model to your targets. This approach has significant benefits over both ETL and ELT, not least because it prevents the data flow proliferation that it is commonplace with traditional tools.

Metadata integration and standardisation

Unfortunately, this is not a trend. We believe it should be. With data catalogues proliferating across a variety of technologies, not just those discussed here, there needs to be some common standards for the interchange of that metadata. There is ongoing work (for example, ODPi Egeria) to create such standards, and this is supported by a number of vendors featured in this Market Update. But they are in a minority, the remaining suppliers claim that that they don't support these initiatives because it has never worked in the past and won't work now. The truth is that it won't work now unless a significant number of vendors buy into standards. But they won't do that because they want to lock you into their environment. They are part of the problem not part of the solution.

Knowledge graphs

Finally, we have noted that a few vendors have started to provide Knowledge Graphs as a part of their platform. Indeed, for one of our vendors, a graph database is intrinsic to its solution. More generally, we expect increasing numbers of vendors to offer knowledge graphs in the future. Their advantage is that they allow you to explore metadata assets across your environment, enable advanced automation (AI), efficient user collaboration, and they help to provide more complete and detailed data lineage and impact analysis.

Vendors

As far as the vendors and products considered in this Market Update are concerned, we have focused exclusively on those that are offering at least four of the five requirements we have outlined. In addition, we have made it a pre-condition that suppliers offer some sort of data integration capability through the deployment of ETL, ELT or some combination thereof. Nevertheless, we have included vendors that provide, for example, a data quality tool but perhaps rely on Collibra to provide data governance. As this is a platform review these suppliers have been marked down where this is the case unless these functions are very tightly integrated. The one exception to this ETL rule is the inclusion of the Qlik Data Integration Platform. Its approach is to offer bulk loading initially followed by change data capture. For companies migrating from an on-premises data warehouse to a new supplier in the cloud this seems a sensible option though it might not be appropriate for all use cases and you lose some of the data integration capabilities, such as push-down and B2B integration, that you might get from a traditional tool. On the accompanying Bullseye chart Qlik has been colour-coded separately from other platforms for this reason. In addition, Microsoft's Azure Data Factory, AWS Glue, Matillion, Fivetran, and Alooma (Google) are also included in the Bullseye, not because they offer a complete or even near-complete platform, but because it is clear that many users are adopting these technologies in their move to the cloud. Detailed descriptions of these products is not provided as a part of this Market Update. In due course, Bloor Research will be publishing a companion Market Update to this one, which will be focusing on stand-alone data integration products like these, and which will include more details of relevant products. A third Market Update on stand-alone data quality tools is also forthcoming while a comparative analysis of data governance products has already been published (see *https://www.bloorresearch.com/research/data-governance-july-2020/*). The one major company that has been omitted is SAS because, although it has substantial existing data management capabilities, the company is in the process of significantly updating its offering and its timescales mean that it would be unfair to position it based on current capabilities and too soon to do so on what is coming. While SAP has been ranked for the purposes of this evaluation, there is no detailed description of the company's capabilities, as SAP did not fully – it did partially – cooperate with our research.

In terms of vendor movement, there have been several recent changes. Firstly, there was the acquisition of Waterline by Hitachi Vantara, without which the company would not have merited inclusion in this report. Secondly, Syncsort acquired the data and software business of Pitney Bowes, and then changed the company name to Precisely. And finally, Vector Capital, the private equity firm, has announced that it will be acquiring MarkLogic. This is expected to be completed by the end of 2020.

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 platform-based solution. Hopefully, the increased availability of managed services and consumption-based pricing will see off the remnants of users that still think that hand coding is a good idea.

Finally, bearing in mind that the various products covered in this Market Update have, at least in some cases, quite different capabilities, we provide a synopsis of these in the following table.

Vendors

	Movement	Quality	Governance	Privacy	Catalogue	Cloud-native	Other
Ab Initio							
AWS Glue							
Ataccama							MDM
Azure Data Factory		Data Prep					
Cloudera		Partner					Various
Fivetran	ELT / CDC						
Google Alooma		OpenRefine					
Hitachi Vantara		Partner					IoT, analytics
IBM Cloud Pak for Data							Various
Informatica							Various
IRI Voracity					Classify		Various
MarkLogic Data Hub					Mostly equivalent		Graph
Matillion	ELT / CDC						
Oracle							Various
Precisely						Planned	Location
Qlik	CDC	Data Prep					DWA, analytics
SAP				Big ID			Various
SAS				Planned	Planned	Planned	Various
Solix Technologies							Archive & Retire
Talend							Data Prep

KEY

Green – Yes

Orange – Some

Red – No (and we are not aware of any specific partnerships)

Metrics

The ideal solution is a broad, unified, automated, integrated, and interconnected platform for data management. Ideally, this should be built on a cloud-native (see discussion above) API-driven architecture. Not only does it contain products and services that, between them, provide all of the essential capabilities we've identified above, it does so in such a way that it is greater than the sum of its parts, by providing additional layers of connectivity, shared metadata, security and so on that elevates the entire platform and creates significant additional value. Needless to say, no existing product meets this ideal – that is, after all, what makes it an ideal – but the vast majority are, in effect, an attempt to approximate it. Some get closer to the ideal than others. As a rule, they all strive to offer high connectivity (with third-party products and environments), high interoperability (offering additional value to their own products by allowing them to integrate with each other), great breadth (via the selection of products available) and extensive automation (for instance, using embedded AI and machine learning). To a very real extent, the degree to which vendor platforms differ is the extent to which they can offer these qualities.

In this context, it is worth commenting that the highest scoring products in this report are generally those with the most advanced cloud-native architectures and the greatest degree of automation through, at least in part, the use of machine learning. Some vendors have been caught out by this and have been late in adopting these technologies; and this also applies to the implementation of data catalogues, whereby some suppliers are further behind the curve than others. Thus this report comes at unfortunate time for those companies that have been slower off the mark than their competitors.

To be specific, scalable, managed and cloud-native deployment to AWS, Microsoft Azure, and Google Cloud should be taken as standard (or in other words, table stakes). Some vendors offer support for multiple clouds (their own, frequently; sometimes other less-used clouds, such as IBM's and Oracle's offerings). A driving motivation for relevant platforms generally is to facilitate increased (and increasing) adoption of the cloud, so this should not come as a surprise. An additional, and more specific, emphasis common to several vendors is the desire to deliver on the agility promised by the cloud in an enterprise-scale setting.

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.

- **Data Movement.** The traditional approach to creating an enterprise data warehouse, the idea here is that you use some combination of extraction (from source systems, typically transactional and operational environments), transformation (because you want the data to be in a consistent format) and loading (into the target system) in order to both create and maintain your cloud data lake or warehouse. These operations may be performed in any order with the historic norm being ETL, though ELT has become popular, especially when loading data into data lakes but also with respect to real-time data that may be streamed into the environment using technologies such as Kafka or Flink. A model-driven approach to data movement is also a possibility. Even restricting discussion to ETL/ELT, the choice is not clear cut: some data integration tools support push-down transformation so that you can perform relevant transformations within the source or target systems, while streaming technologies often have some level of transformational capabilities built into their platform. In this category, the varieties of methodology available within each product (which also include bulk loading, change data capture and so on) is a significant factor in its score, as are the depth of features on offer (the aforementioned push-down transformation being a good example). There is an overlap with Platform Integration (see below) with respect to connectivity.
- **Data Quality.** The ability to ensure the data you are processing is complete, consistent, accurate and fit for purpose. In short, that it is trustworthy. This makes it much easier – or arguably, possible at all – to gather consistently accurate and trustworthy analytics from your data, which in turns allows you to make decisions based on your analytics soundly and with confidence. In all likelihood, you will want to profile and cleanse your data – thus ensuring its quality – as it is ingested into your environment, but in addition, it's a good idea

to systematically profile and monitor your data over time in order to make sure that this level of quality is maintained. There are, broadly speaking, two parts to this: first, you need to be able to test (profile) your data for quality, monitor said quality over all of your data, and hence know when (and where) your data quality is lacking; and second, you need to be able to repair (cleanse) poor quality data when you find it. The first of these is, generally speaking, the more competitive area of the two, and hence will usually be the more significant determinant for this metric. It should be noted that some vendors in this space rely on landing data into your data lake and then using the profiling and transformation capabilities of their data preparation tools for data quality purposes. For some initial exploratory use cases this may be adequate. However, we do not regard this as a best practice approach having the same breadth of functionality and applicability as offering data quality tools per se.

- **Data Cataloguing.** Being able to centrally manage, explore, search and access all of your data (and metadata) using a common interface is a powerful capability: hence, the data catalogue. Metadata management, data discovery, and (visual) data lineage are all core capabilities, often integrated with a business glossary. Automated management capabilities, such as the automatic tagging of data and linking to business terms, or the implementation of machine learning driven data discovery and dataset recommendations, are also particularly appreciated here, as is interoperability (after all, what use is a universal catalogue if it's not appreciably universal?). Ease of use and collaboration are also important, the latter often facilitated by shareable searches, commenting functionality, user reviews, and so on. Of note (although to an extent this falls under integration) is the ability to push and pull data from other catalogues, and hence integrate with them, perhaps even forming a *"catalogue of catalogues"*. Support for ODPi Egeria or another metadata standards-based approach may also be a factor in achieving this.

- **Data Governance.** In its broadest sense, data governance can cover a number of different areas, including data quality, data cataloguing, data privacy, and policy management. For the purposes of the report, we use it primarily to talk about the last of these: how one defines, manages and (in an ideal world, automatically) enforces organisational policies for data quality and compliance mandates. Compliance encompasses a variety of regulations, most prominently including GDPR and CCPA. A product that scores well in this category will almost certainly have built-in provisions for a number of mandates, but will also be extensible, validate the implementation of policies, and be able to change over time as new regulations emerge or existing regulations are altered or overwritten. This also applies to internal, organisational policies, and note that while compliance mandates are obviously important, they are not the be-all and end-all for policy management. We should add that data governance in general is becoming more focused on business outcomes and data democratisation and features that facilitate this will be welcome.

- **Data Privacy.** While data governance covers data quality and regulatory compliance at the policy level, data privacy covers it at the data level. In other words, governance determines what your policies are and whether they're enforced, but data privacy determines how they are implemented, at least with respect to personal data. This will inevitably involve finding, protecting, and securing your sensitive data, usually using some form of data masking as well as sensitive data discovery, role management, and role-based access. From a data masking perspective this will need to be dynamic and multiple masking algorithms should be supported, not least those that allow consistent masking and referential integrity. It will be useful if there are multiple ways of implementing data masking (whether in situ, using a proxy server or via APIs, for example). This metric also covers data archival and retention – and, in particular, the ability to purge personal data on request – as well as support for DSARs (Data Subject Access Requests). Important factors here include breadth, ease of use and implementation (with the least intrusive solutions being the more highly scored).

- **Architecture.** There are two aspect to this. Firstly, there is interoperability. While integration concerns how a given platform connects with the rest of your system, interoperability cares about how well the different products within a platform connect with each other. In essence, this is the *'value-add'*: what separates a platform from a mere collection of products. This can easily tie into automation as well, and in fact a common way for a platform to add value is to automate the myriad of tasks related to data integration and management. Equally important is the sharing of data and metadata of the products within the platform, which can be accomplished by providing a shared metadata layer that every product within the platform can draw from.

  Secondly, architecture relates to the way in which the platform is deployed, in particular whether it is cloud-native (as discussed previously) but also to what extent it supports traditional virtues such as performance, security and so on, as well as self-service, collaboration and ease of use, each of which also overlap with the previous five metrics.

- **Automation.** Self-evidently important, the automated capabilities present in a data management platform can be a very significant differentiator, and in many ways is a matter of *"the more, the better"*. The most highly automated platforms 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. The provision of (automatically generated) knowledge graphs that support the unification and management of metadata is also a factor in this metric, as is a focus on the explainability of any machine learning algorithms being deployed.

- **Platform Integration.** Being able to fit into an existing ecosystem is an important property for any data product and this not only applies to integration across the products within the data management platform (see Architecture), but also to integration with the outside world. In some cases, this may mean deep integration between the platform provider and the third-party (for example, close integration with Collibra) or it may be through support of open APIs.

  The other aspect of third-part integration is connectivity. Bear in mind that there are more than 350 databases available on the market, that the number of applications is also measured in hundreds, if not thousands, and that there are a whole host of sensors and other edge devices that might need to be supported as sources, and you can see that it is impossible for any vendor to cover all of these with native connectors. Nevertheless, support for leading products in the various categories should be expected and we are disappointed that a number of the newer products are very limited in this respect, having to rely on generic connectivity options such as ODBC/JDBC, which will not perform as well as native connectors. On the other hand, leading vendors that offer hundreds of native connectors are still only scratching the surface. The provision of a software development kit (SDK) so that new connectors can be quickly developed, will be desirable.

Conclusion

The facilities included within (Cloud) Data Management Platforms are fundamental to the successful implementation of cloud-based data warehouse and lakes, as well as to other data integration use cases. Compared to partial offerings and home-grown cloud stacks, they are almost always more tightly knit and frequently far more comprehensive. As such, every platform included in this report comes highly recommended, at the very least when compared to assembling your own solution, or to going without a solution entirely. Even if you don't operate in the cloud, and don't intend to in the near future, the platforms discussed here may still be worth your time and consideration: the breadth of data management capabilities they offer within a single location is simply too significant to overlook.

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**