

Data Lake Management

Market basics

Data Lake Management (DLM) encompasses all the functions that are necessary to load, manage, govern and secure one or more data lakes within the context (where appropriate) of a broader data ecosystem (in other words, also incorporating data warehouses and other more traditional environments). According to the Gartner Group: *“through 2018, 90% of deployed data lakes will be useless as they are overwhelmed with information assets captured for uncertain use cases.”* In our view, this *“ain’t necessarily so”* and DLM is the key to ensuring that data lakes do not become data swamps. On the contrary, DLM should enable and facilitate business analysts, data scientists and others that wish to explore the data in their lake(s).

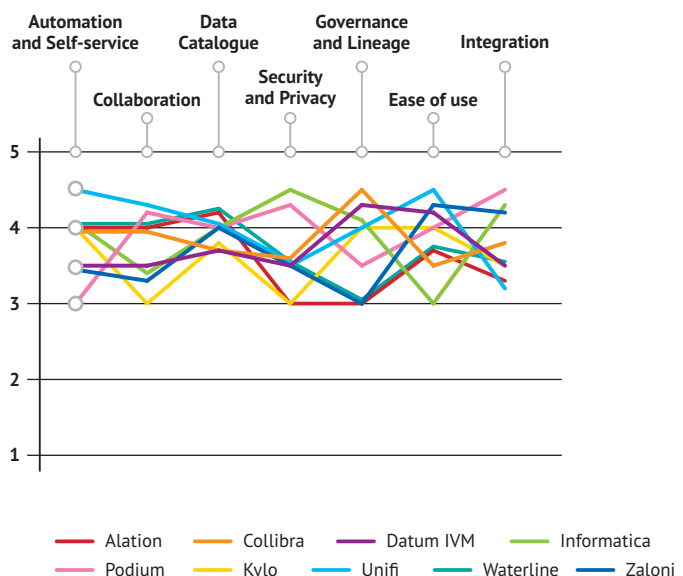
In practical terms, there are a couple of different approaches to DLM available in the market. All vendors have solutions that focus on the creation of an indexed, searchable data catalogue that supports self-service. In addition, features such as metadata management, ingestion management, the identification of sensitive data, and access security should be considered mandatory. Where vendors differ are in:

1. Whether they also provide data preparation capabilities (and, even if they have them, whether these are actively marketed). This is a point worth discussing. Our view is that while data cataloguing and data preparation are both required to enable a self-service environment only the former is strictly necessary for data lake management and governance. On the other hand, data profiling (often included within data preparation) will be necessary to identify sensitive information, and it will be advantageous to be able to profile data on ingestion.
2. Whether they are broader data governance solutions that potentially apply across the enterprise rather than just applying to data lakes. Note that this does not preclude the provision of policy-based capabilities within more specialised DLM products but, where this is the case, these tend to be data lake specific. In particular, it will be important to have the ability to define policies or rules around sensitive data and how that should be treated.

There is also one vendor (currently) that offers all of these capabilities, albeit in multiple products.

Market trends

In 2014, we researched the market for data preparation tools and produced a Market Update reporting on the results in 2015. Later that year we returned to this subject because of the number of new data preparation tools in the market, and we produced a revised report – published in 2016 – that also extended our analysis to tools for discovering metadata within data lakes and creating searchable catalogues against that metadata. We did this because we felt that knowing what data was in your lake was a necessary precursor to preparing data for analysis. However, the world has moved on. There are now, probably, more than fifty different products that provide data preparation – basically all the analytics suppliers now offer it, as well as pure play providers and data management vendors – compared to just eight in 2014. Similarly, while you could count the number of cataloguing vendors on the fingers of one hand when we first started investigating that market, there are now well over thirty products that incorporate catalogues, notably from analytics vendors and data governance suppliers.



Compared to these figures the number of vendors in the DLM space is relatively small. While business intelligence products may now have data catalogues built-in, they certainly don't have the governance and other features that we would expect of a DLM offering. In practice, therefore, the DLM space is dominated by vendors that started with data catalogues and have expanded from there. However, while most of these are start-ups that is not always the case, and some established data management vendors have also taken this approach.

In terms of market development, we see four major trends. Firstly, we expect at least some data preparation vendors to develop their products into full-blown DLM offerings, and some are already doing so, though we do not consider any of these to be mature enough at present, to be included in this report. Secondly, we anticipate that (policy-oriented) data governance suppliers will increasingly extend their capabilities to support DLM. This is already true to a certain extent, though currently such products tend to be weaker in terms of specific data lake features. Thirdly, we expect more general-purpose data management vendors to extend their existing platforms into the DLM arena. Again, this is already happening. Fourthly, we are starting to see vendors introduce template-based and packaged applications for specific scenarios. For example, there are already a couple of vendors offering packaged GDPR (General Data Protection Regulation) solutions. Finally, all of these trends will put pressure on data quality vendors to extend their solutions so that they can handle DLM requirements. This will lead to acquisitions and consolidation across the market as companies look to offer more complete and comprehensive solutions.

In effect, pure players will get squeezed, as they always do. On the one hand from data management companies moving into their space and, on the other hand, by data governance vendors, not to mention companies that do both. The question will be whether one or more of the pure players can develop enough momentum (through a combination of technical advantages and marketing skill) to establish themselves as a market leader.

One final point worth discussing is the role of open source in DLM. There is a recognisable trend in some companies to adopt an "everything open source" strategy so it is worth asking whether there is an open source DLM solution? The short answer is no. There are partial solutions that are open source and there are proprietary vendors who market their products as open source, but in the true sense of the term: no or, at least, not yet.

Vendors

In this section, we provide details of the various vendors and products targeting DLM. Briefly, these fall into four groups:

- **Pure-plays that may or may not offer data preparation.** Those that do include Podium Data, Teradata (Kylo), UNIFI and Zalon. Alation does also but it does not emphasise this and instead recommends one of its partner solutions from companies such as Paxata and Trifacta. It is worth commenting that Alation is also a partner of, and is resold by, Teradata. The claim is that Alation and Kylo have complementary capabilities and, certainly, there are joint customers using both in conjunction. However, we do not see how this position can be sustained in the longer term: most users would prefer to have one tool than two. Finally, Waterline does not offer data preparation and relies on partnerships with third parties for this purpose. We considered including both Tamr and Alteryx in this category, the former because it has a data catalogue and the latter because of the recent release of Alteryx Connect. However, we have concluded that neither company has sufficient depth of solution to be truly classed as a DLM solution. In the case of Alteryx, in particular, its product is based on its acquisition of Semanta, which is Windows oriented rather than focused on Hadoop.
- **Data governance vendors.** We have included two data governance vendors – Collibra and DATUM – in this report as exemplars of policy-oriented data governance providers that include capabilities for DLM. The most likely reason for selecting one of these suppliers for DLM is because you want a consistent enterprise-wide policy-oriented governance solution, so the emphasis will be on data governance per se rather than data lakes. As our next Market Update will be on data governance we have left detailed consideration of all such vendors to that report.
- **Data management providers.** By this we mean companies whose offerings span, DLM, data preparation and data governance. At present, the only company in this class, and in this report, is Informatica. As a part of the research for this paper we looked in detail at both Talend and Pitney Bowes but neither of these currently has a sufficient breadth of capability to be considered as a genuine DLM contender. In this category, it is also worth mentioning that IBM has announced Unified Governance, which will be based on Apache Atlas, so we expect IBM to be a potential provider at some point in the future.

Conclusion

We have evaluated products along seven axes:

- **Data catalogue** – this is central to data lake management. As might be expected, Alation and Waterline, which specialise in the creation and use of catalogues, are leaders in this area.
- **Automation** and the extent to which products lend themselves to self-service. A number of products are strong in this area, but we especially like UNIFI's capabilities.
- **Collaboration**, including support for things like crowd-sourcing. Podium Data and UNIFI are strong in this area though they are only the first among several.
- **Security**, including the discovery of sensitive data, the ability to set security policies, data masking, access control, and integration with relevant open source projects and authentication standards. While vendors are generally strong when it comes to discovering sensitive data, they mostly rely on third parties for anonymisation. The exceptions, which rate most highly in this category, are Informatica and Podium Data.
- **Governance**, including data lineage and the ability to remove redundant, out-of-date and trivial (ROT) data. As might be expected it is the suppliers that have more general data governance capabilities that lead here, notably Collibra, DATUM and Informatica.
- **Ease of use**, including packaged applications and templates. In the case of Zaloni this includes its "Data Lake in a Box" offering. The vendors we like here are UNIFI, Zaloni and DATUM though several others also have strong ease of use capabilities.
- **Integration**, including the ability to ingest legacy data from sources such as VSAM or COBOL copybooks, as well as more modern sources such as the ingestion of streaming data. Apart from Informatica, which you would expect to be strong in this area, the other vendors with impressive integration capabilities include Podium Data and Zaloni.

As can be seen, all the products are leading in one category or another. As a result, and this is unusual for a Market Update – in fact this has never happened before – this report does not include a Bullseye Chart. The reason for this is that all the products we have reviewed are all eminently capable of supporting the management and governance of data lakes. Of course, some products are better at some elements of this than others but overall, once we had completed our calculations, there was so little difference between the products that we feel that it would be unfair to separate them, especially when you bear in mind that there is always some element of subjectivity in making these evaluations.

For those familiar with Bullseye charts and who really want scores: all the vendors would be clustered around a score of 4.0 (actually, 3.8 to 4.2 out of a total of 5). Further, we would have to argue that all the suppliers considered here are innovators although, obviously, some are larger companies and more well-known than others.

Waterline Data Catalog


Waterline Data Catalog is a complete solution for data discovery, data tagging, curation and cataloging for the purposes of enabling self-service data exploration, rationalisation and compliance (for instance, with GDPR). It features data profiling, data categorisation, and global search based on Apache Solr, each of which works across all of your systems, including both data lakes (Apache Hadoop, with support for Apache Spark) and relational databases (via a plugin architecture). Lineage can be imported from another source (for instance, Cloudera Navigator or Apache Atlas) or derived directly from your data (and corrected manually if needed).

Waterline recognises that the people in your organisation, know a great deal about your data. Consequently, Waterline seeks to extract that knowledge and formalise it so that it can be accessed freely across your company. This has led to an emphasis on collaboration and crowdsourcing that pervades the product. For instance, users are encouraged to leave ratings and reviews on data sources, and data stewards in particular can annotate data sources in order to guide and help other users. Combined with Waterline's data profiling capabilities, this allows you to analyse your data using both objective and subjective information.

Waterline also has sophisticated data matching capabilities that leverage machine learning to automatically suggest business terms – known as 'tags' – for fields within your data sources. This is done by examining the data itself, rather than simply the field name. Users with appropriate authority can accept or reject these suggestions. If accepted, they are added to the field as a custom attribute. What's more, due to the machine learning, these suggestions will become more accurate over time. The tags themselves are stored within your system, each associated with a particular domain. Built-in tags exist in a default domain, while additional prefab domains, that contain common tags used in a particular space, such as 'GDPR' or 'Retail', are available. Existing business glossaries or taxonomies can be imported and additional terms and domains can be created manually. Waterline also features automated tag-based data access control. Combined with data matching, this makes it very easy to protect even large amounts of sensitive data as it flows into your data lake. Integration with Apache Ranger and Cloudera Sentry is provided and the company partners with data preparation vendors such as Trifacta and Paxata, as well as with Privitar for securing sensitive data.

Strengths

- Automated data matching makes it much quicker to tag your data. It also makes it easier to find and not miss fields that need to be tagged. This is



Waterline Data
 201 San Antonio Circle Suite 260
 Mountain View CA 94040
www.waterlinedata.com

particularly important if you are tagging sensitive data and cannot afford to overlook such a field (for instance, in order to comply with GDPR).

- Waterline's emphasis on automated discovery and categorisation combined with human collaboration makes it easy to disseminate important information relating to your data, which can make working with your data sources much simpler.
- Waterline provides a wide variety of APIs for integrating its data catalogue into existing data work flows. This is something many competitors claim to do, but Waterline has customers who specifically call out this capability as a differentiator.

Threats

- Although Waterline offers a complete data discovery and cataloging solution, it does not extend into data preparation. Whether this is threat or a strength depends on your point of view. If you already have a data preparation solution or are looking to purchase a "best of breed" technology stack, then it is a strength.

Summary

Waterline is a specialist data lake management vendor that partners and integrates with third party data preparation suppliers. As might be expected of such a vendor, its data cataloguing is especially strong and its self-service, automation and collaborative capabilities are also excellent.

