REPORT REPRINT

Sink or swim? Governance and preparation are key to a functional 'data lake'

MATT ASLETT

3 MARCH, 2016

It may be an oversimplification, but the term 'data lake' is here to stay. Data governance and self-service preparation are the foundations that will will turn the data lake from concept to reality.

 |

Two years ago, we argued that 'data treatment plant' was a more suitable term than 'data lake' for the use of Hadoop to ingest data from multiple sources into a unified data platform serving multiple applications. We were aware that it wasn't a battle we were ever likely to win, and the term 'data lake' has become near ubiquitous since then. We stand by our argument that the need for data to be filtered, processed, treated and managed to make it suitable for multiple analytics use cases is critical to delivering value from the data lake. Two key trends are now driving the creation of these environments: data governance and self-service preparation.

THE 451 TAKE

As we recommended in our 2016 Trends in Data Platforms and Analytics report, "enterprises should seriously consider the data governance and management requirements before embarking on data lake projects to ensure that the functionality is available to turn the concept into reality." The emergence of a variety of data lake management products, combined with an increased focus on governance within the Hadoop community, means that it is now considerably easier for enterprises to establish the data governance and management fundamentals on which to create their data lakes, while self-service data preparation provides the user interface for reducing the time taken to analyze the data and extract true value from the data lake. The combination of self-service data preparation, Hadoop and data governance is still evolving and maturing, however, and provides multiple opportunities for partnerships, mergers and acquisitions.

CONTEXT

One of the prime advantages of Apache Hadoop over existing analytic database technologies is its schema-on-read approach to data processing, meaning that users can ingest data into Hadoop without first having to define how they are going to analyze it – in stark contrast to the schema-on-write approach used by traditional analytic databases, in which the schema and data model is defined in advance in order to deliver high performance.

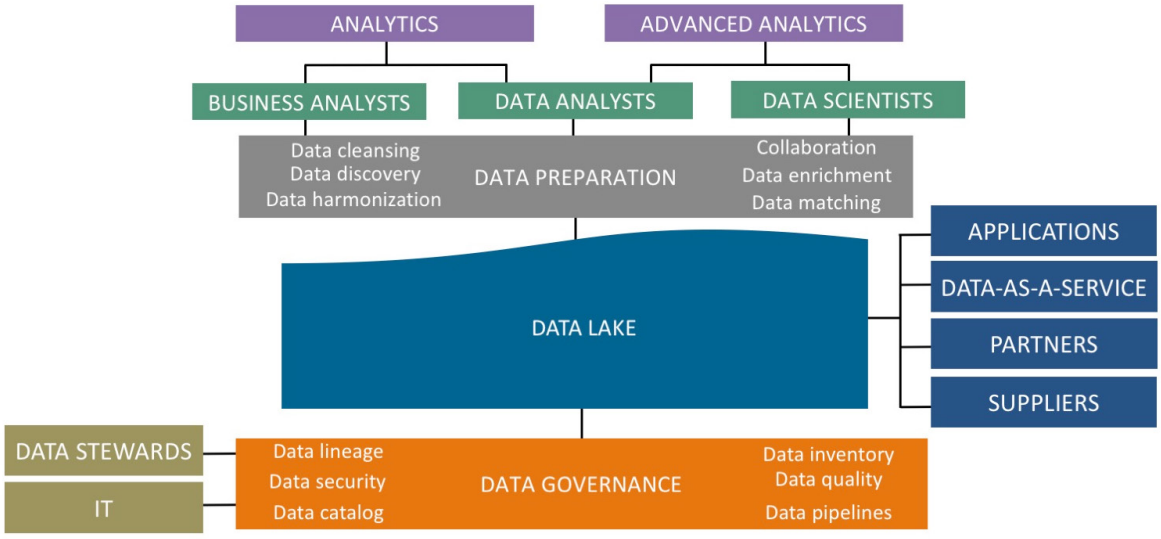
The flexibility of schema-on-read has encouraged many organizations to turn to Hadoop as a landing zone for data from multiple applications – particularly those generating unstructured data – and the desire to use Hadoop's flexibility (driven by the emergence of YARN to enable Hadoop to run multiple workloads) to treat it as a single logical unified data platform serving multiple users and multiple applications.

The term 'data lake' is credited to Pentaho's founder and CTO, James Dixon, who first used it in a 2010 blog post in which he described it as a large body of water that is fed from various source streams and that can be accessed by multiple users for multiple purposes. Dixon compared the data lake to a data mart, which he said could be considered the equivalent of bottled water: cleansed and packaged for easy consumption.

In early 2014 we argued that what 'data lake' failed to address, as an analogy, was how multiple users would access that data for multiple purposes. Offering up an alternative analogy – the 'data treatment plant' – we argued that industrial-scale processes were required to make data acceptable for a desired end use and the multiple methods for accessing and processing data.

While the term 'data lake' has taken off, greater emphasis is now on those industrial-scale processes that are required to turn the data lake concept from theory to reality. There are two key drivers for the focus on these industrial-scale processes: data governance and self-service data preparation. Below, we will look at how each of these enables the data lake and complements each other.

DATA GOVERNANCE, DATA PREPARATION AND THE DATA LAKE



DATA GOVERNANCE

While Hadoop offers a more flexible schema-on-read approach to analytics, it is clear that it is also not a free-for-all. In order for data to be landed from multiple sources and made available to multiple users for multiple purposes it has become clear that some key concepts needed to be addressed, including data ingestion, data discovery, data inventory, data enrichment, data quality, metadata management, data lineage and data security.

The data security angle is clear – enterprises want to make sure that they are able to restrict access to data to users for which it would be unsuitable. Beyond that, however, data governance in relation to the data lake is less about restricting access to data as it is about enabling access to the right data. This is especially important in Hadoop since it is made up of multiple projects, many of which have their own data governance capabilities, but lacks mature cross-project data governance capabilities.

At the very least, a data catalog is required so that users can create an inventory of exactly what data is in the environment in the first place. This data catalog (and metadata catalog) can then be used by analysts to discover data sets for analysis while data stewards will be looking to apply data quality and master data management tools to ensure that the data is fit for analysis. Data lineage is also important in this regard in enabling analysts and data stewards to understand where the data came from, and what transformations may have been made to it already. Data lineage and data governance is also vital to compliance projects, which are traditionally managed by IT, or the data steward in larger enterprises.

In order to ensure that multiple workloads can run efficiently against a single Hadoop cluster, the ability to create and manage data pipelines and monitor and manage workflows is also important (and addressed further in this related 451 Research Spotlight report).

There are a variety of vendors targeting the role that data governance has to play in delivering a managed data lake, including startups such as Waterline Data with its Data Inventory offering, Zaloni with its Bedrock Data Management Platform and Alation with its Data Catalog. Podium Data is another vendor in this space with its eponymous data lake management platform.

Established data management vendors are also positioning for data lake management, including Informatica with Big Data Management, Talend with its data fabric, Hitachi Data Systems' Pentaho with its focus on the analytic data pipeline, and SnapLogic with its growing self-service capabilities, especially in relation to Spark-based data pipelines. Global IDs is another example of a vendor looking to support governance, profiling and other data management arenas for Hadoop-based data lakes.

The major Hadoop distributors are also increasingly tackling the issue of data governance. Cloudera has offered its Navigator product since early 2013 while Hortonworks is backing Apache Falcon as a framework to simplify data pipeline processing and management and in early 2015 teamed up with the likes of Aetna, Merck & Co, Target Corp and SAS Institute to create launch an initiative to improve data governance in Apache Hadoop. One of the results is Apache Atlas, which includes a metadata service and UI for searching for metadata and lineage, as well as integration with Hive.

These Hadoop distributor initiatives are for the most part complementary to the specialist governance products described above, and are designed to enable the ecosystem by providing baseline functionality in Hadoop itself. For example, Alation is certified with Cloudera Enterprise, while Waterline Data Inventory is certified with Cloudera, Hortonworks, MapR and Pivotal and Zaloni has partnered with Cloudera, Hortonworks, MapR, Pivotal and IBM.

DATA PREPARATION

As noted above, one of the primary advantages of creating a catalog of data in Hadoop is that it can be used by analysts, data scientists and business users to discover and prepare data sets for analysis. This is where self-service data preparation comes into play as a driver and enabler for generating value from data lake deployments by reducing the time taken to prepare data for analysis.

We have previously covered the rise of self-service data preparation as a means to reduce the burden on IT to prepare data for end users, and in doing so reduce the time taken for users to discover, integrate, cleanse and enrich data to make it suitable for analysis.

Initially driven by a range of specialist startups including Trifacta, Tamr and Paxata, self-service data discovery and preparation has now been embraced by all the major data management vendors as well, including Oracle, IBM, SAS Institute, Informatica, Talend and HDS's Pentaho, as well as a variety of Hadoop-focused analytics players including Alteryx, Platfora, Datameer, ClearStory and Datawatch.

Self-service data preparation is not unique to Hadoop-based data lakes in any way – many of the products above can be applied to traditional analytic databases – but self-service data can be seen as a key enabler in reducing the time taken to identify and combine data from multiple sources for analysis and delivering value from a data lake deployment.

There is an element of overlap with the data catalog and governance offerings detailed above. Tamr's self-service data preparation offering is based on its data cataloging capability, combined with attribute mapping and record matching, for example, while Alation's Data Catalog is targeted primarily at non-technical business users with a search-based interface for discovering and preparing data.

UNIFi Software is another new player with a catalog search approach and capabilities for preparing and integrating data from multiple sources, while Harte-Hanks' Trillium Software used UNIFi's software as the basis of its Trillium Prepare and Trillium Refine products, which offer data preparation and data preparation plus data quality capabilities, respectively. Additionally, Podium Data offers a combination of self-service access based on a searchable data dictionary with underlying security and data governance capabilities.

The incumbent data management vendors, including Informatica, Talend and Hitachi's Pentaho, clearly see self-service data preparation as part of a suite of products that also provide data stewards with the capabilities to manage, curate and govern data.

Even the self-service data preparation specialists, while aimed squarely at removing the barriers to insight for business analysts, data analysts and data scientists, have introduced data governance capabilities to meet the needs of the IT organization.

For example, Paxata added governance and workload management advances in its fall 2015 release that were enhanced in its more recent winter 2015 release, while Trifacta added data governance, lineage and other features required by IT to manage Hadoop in version 3 of its Wrangler Enterprise product.

ClearStory Data is another example. While the company might be best known for its data-harmonization, analysis and collaboration software for business and data analysts, under the covers it relies on automated data profiling, data inferencing, data lineage and data governance capabilities that are also proving attractive to customers interested in data lake deployments.

Data lake governance and self-service data preparation are potentially complementary areas of focus, however, as illustrated by Zaloni's launch of its Mica self-service data preparation offering as a complement to the Bedrock Data Management Platform, as well as Waterline's partnership with Trifacta. Indeed, we would anticipate the potential for merger acquisition activity to drive greater convergence.

Similarly, while the Hadoop distributors have – so far at least – decided to stay out of the self-service data preparation space, preferring to partner with the likes of Trifacta, Paxata and Tamr, we think data preparation, along with data governance, could provide merger and acquisition opportunities for Hadoop distributors in 2016 and beyond.