On the Radar: Unifi Software provides a Swiss army knife for data lake integration and governance

Summary

Catalyst

The need to make data lakes transparent and the emergence of machine-learning capabilities have combined to spawn a vast new marketplace for data preparation and curation tools. These capabilities have become available, either as extensions of data integration, BI, or data science platforms, or through standalone tools that are independent of the analytics tool. But as these capabilities have become broadly available, some providers are taking the next logical step in extending their solutions to data governance. Among the independents with the broadest set of capabilities is Unifi Software, which offers a Hadoop-native data integration tool that, uniquely, addresses both the needs of IT data engineers and business analysts. Unifi's greatest challenge is communicating the value proposition of a broad-based platform for integrating, curating, and managing the governance of big data in a developing market for data lake governance that is full of point solutions.

Key messages

- Unifi Software offers one of the broadest suites for preparing and curating data among independent providers, combining data preparation and cataloging with managing governance workflows.
- Unifi's opportunity is extending its machine-learning capabilities to make curation and governance of big data a fully guided experience.
- Unifi competes in a landscape that is populated, on one end, by point self-service tools that are either standalone or extensions of analytic suites and, on the other end, classic data integration providers that are new to self-service.

Ovum view

Unifi Software offers one of the most complete suites for preparing, curating, and governing big data among independent players and is best suited for organizations with multiple analytic tools, which require integration capabilities that are tool independent. Its challenge is getting heard above the growing noise in this space and leveraging machine learning to perform related tasks, such as optimizing data ingestion and managing master data.

Recommendations for enterprises

Why put Unifi Software on your radar?

Consider Unifi Software if your organization's requirement is for a broad-based self-service approach to integrating, preparing, and managing the governance workflow for big data that is not tied to any specific analytic or BI query tool.

Highlights

Background

In the Ovum report, *Developing a Strategy for Data Lake Governance*, we identified the need for IT and end users alike to know the inventory of the data lake. In contrast to traditional data-warehousing scenarios, the data lake would require shared responsibility between IT *and* business analysts/end users for integrating data and populating the data store. That is because of the nature of the data lake, which is to support exploratory analytics; IT would not have the bandwidth to generate a new data set each time the business wants a different view.

With the emergence of machine learning as a practical approach for helping practitioners perform data transformation and integration, a new market blossomed for supporting discovery, transformation, and integration of data for the data lake, addressing data sources that are far more varied in content and structure compared to classic data warehouses. Thanks to machine learning, this new generation of tools has extended self-service upstream from data visualization to data preparation and curation.

Unifi Software, which was started by some of the founders of Greenplum and Alpine Data, offers a broad-based tool spanning from data discovery to profiling, preparation, cataloging (which does not always come as part of a data preparation tool), and managing the workflow of governance. Working with structured or variably structured data sources, Unifi automatically harvests metadata as data sets are identified; this metadata populates the catalog, which can subsequently be enriched by business users (e.g., with annotations or ratings). It includes the profiling capabilities that typically accompany data preparation tools. Its patent-pending transformation, "OneParse," scans the data to identify its structure; inserts missing values and detects quality issues by row; and triggers Hadoop processes (Hive or Spark) to perform transformations and cleansing. Unifi incorporates machine learning for recommending joins based on detection of common keys across multiple data sets; identifying KPIs based on value ranges in specific columns; and selects the appropriate transformation function (from a library of more than 500 that come with the tool). In turn, it helps data stewards manage the workflow of governing big data by providing capabilities to tag and automatically notify data engineers of the tasks that must be performed to provide access to the data for business analysts.

Under the covers, Unifi runs on Hadoop and has a cost-based optimizer that works with YARN to designate the optimal approach (e.g., using Hive and/or Spark) for the operation. With adapters, it provides data in native format for BI and visualization tools such as Tableau, Qlik, and Looker. And business users can find data through natural language text search. Unlike most data preparation tools, Unifi does *not* use a spreadsheet interface; instead, it provides a drag-and-drop front end where users manipulate columns while performing joins or splits of tables.

Unifi's strength (and core positioning) is that it is far more broad-based than most so-called data preparation tools. It allows data engineers to turn transformations into one-click recipes that can be reused. Like some data preparation tools, Unifi has additional cataloging capabilities, although its capabilities are not necessarily the same as those offered by point tools such as Alation (which heavily targets query optimization). But more importantly, Unifi is tackling the management of the data governance workflow. It starts with automatic tagging of sensitive data, from which data stewards can then instruct data engineers to take the appropriate action to protect the data through encryption, masking, or redaction. Unifi provides dashboards that enable data stewards to track that their data
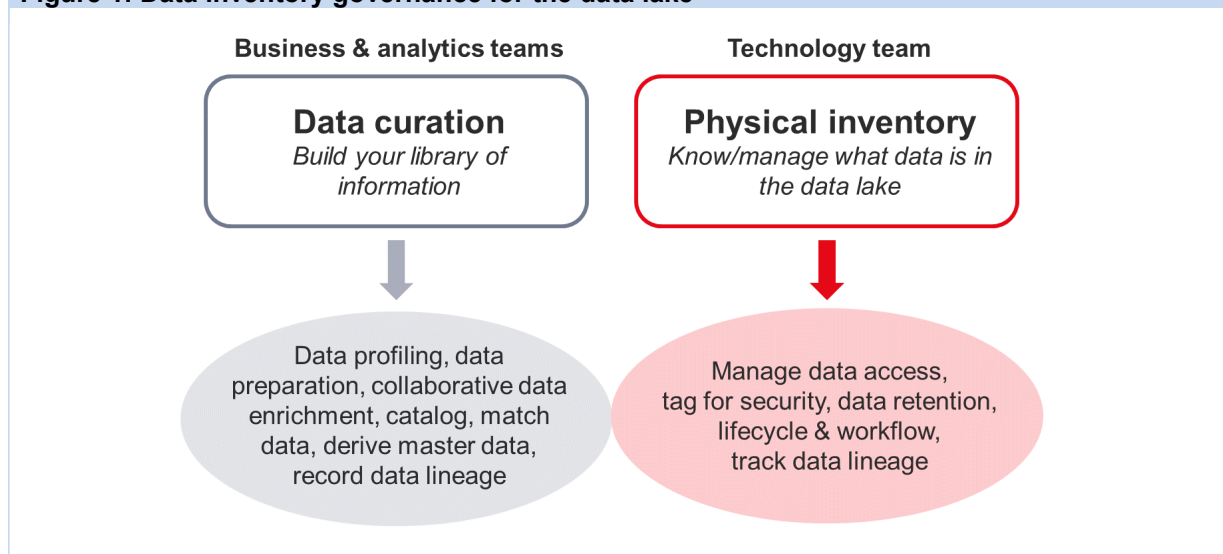
governance instructions have been carried out, while tracking related areas such as incidence of improper or failed data access attempts.

Current position

The company, which roughly doubled its financing to $32m with a B round of venture financing back in March, competes against independents such as Paxata and Trifacta that have drawn roughly twice the funding. An early foray into a strategic partnership with data quality tool provider Trillium was aborted after the company was acquired by Syncsort in a private equity-financed deal last year.

When viewed as a tool for governing data lakes (see Figure 1), Unifi addresses the data inventory tier and manages the governance workflow – it does not (yet) fully automate governance. It addresses most (but not yet all) end-user curation functions and some IT physical inventory functions (managing data access and protection). It also provides a capability for resource management of its jobs on Hadoop clusters. Unifi's challenge is that as it expands its role in managing the inventory and curation of data in the data lake, it must clarify where it fits along with those of partners such as Hadoop platform providers – many of whom also track data lineage, perform cataloging, and protect and manage access control to data. We believe that Unifi's biggest opportunities will lie with cloud providers such as Amazon and Google, because their managed Hadoop services (EMR and Cloud Dataproc, respectively) do *not* wade into these areas.

Figure 1: Data inventory governance for the data lake

Source: Ovum

The key draw for Unifi customers is a tool that plays a "Switzerland" role that is not exclusively tied to any specific analytic tool or database, and that offers a broader set of functions for curating and integrating the inventory inside the data lake. With impressive machine-learning capabilities that are already built into its toolset, Unifi has the opportunity to expand the capability in optimizing the lifecycle of ingesting and curating data in the data lake by applying machine learning to make the curation and transformation of big data a guided experience.

Data sheet

Key facts

Table 1: Data sheet: Unifi

Product name	Unifi	Product classification	Data preparation, cataloging, and management of governance workflow
Version number	2.x	Release date	Every six weeks
Industries covered	All	Geographies covered	North America, EMEA
Relevant company sizes	Midsize–large	Licensing options	Subscription
URL	www.unifisoftware.com	Routes to market	Direct and partners
Company headquarters	San Mateo, California, US	Number of employees	63

Source: Ovum

Appendix

On the Radar

On the Radar is a series of research notes about vendors bringing innovative ideas, products, or business models to their markets. Although On the Radar vendors may not be ready for prime time, they bear watching for their potential impact on markets and could be suitable for certain enterprise and public sector IT organizations.

Further reading

Developing a Strategy for Data Lake Governance, IT0014-003113 (May 2016)

Data Lake Governance: Answering Some Common Questions , IT0014-003114 (May 2016)

Author

Tony Baer, Principal Analyst, Information Management

tony.baer@ovum.com

Ovum Consulting

We hope that this analysis will help you make informed and imaginative business decisions. If you have further requirements, Ovum's consulting team may be able to help you. For more information about Ovum's consulting capabilities, please contact us directly at consulting@ovum.com.

Copyright notice and disclaimer

The contents of this product are protected by international copyright laws, database rights and other intellectual property rights. The owner of these rights is Informa Telecoms and Media Limited, our affiliates or other third party licensors. All product and company names and logos contained within or appearing on this product are the trademarks, service marks or trading names of their respective owners, including Informa Telecoms and Media Limited. This product may not be copied, reproduced, distributed or transmitted in any form or by any means without the prior permission of Informa Telecoms and Media Limited.

Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Informa Telecoms and Media Limited nor any person engaged or employed by Informa Telecoms and Media Limited accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard – readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Informa Telecoms and Media Limited.