The Forrester Wave™: Machine Learning Data Catalogs, Q4 2020

The 10 Providers That Matter Most And How They Stack Up

by Michele Goetz

October 14, 2020

Why Read This Report

In our 39-criterion evaluation of machine learning data catalog providers, we identified the 10 most significant ones — Alation, Alex Solutions, Collibra, data.world, erwin, Hitachi Vantara, IBM, Infogix, Informatica, and Io-Tahoe — and researched, analyzed, and scored them. This report shows how each provider measures up and helps enterprise architecture (EA) professionals select the right one for their needs.

Key Takeaways

Alation, Collibra, Alex Solutions, And IBM Lead The Pack

Forrester's research uncovered a market in which Alation, Collibra, Alex Solutions, and IBM are Leaders; data.world, Informatica, Io-Tahoe, and Hitachi Vantara are Strong Performers; and Infogix and erwin are Contenders.

Collaboration, Lineage, And Data Variety Are Key Differentiators

As metadata and business glossary technology becomes outdated and less effective, improved machine learning will dictate which providers lead the pack. Vendors that can provide scale-out collaboration, offer detailed data lineage, and interpret any type of data will position themselves to successfully deliver contextualized, trusted, accessible data to their customers.

The Forrester Wave™: Machine Learning Data Catalogs, Q4 2020

The 10 Providers That Matter Most And How They Stack Up

by Michele Goetz
with Gene Leganza, Boris Evelson, Jennifer Belissent, PhD, and Robert Perdoni

October 14, 2020

Table Of Contents

Related Research Documents

The Forrester Tech Tide™: Data Management, Q1 2020

Put The Business Back In Your Data Management Business Case

Research Overview: Machine Learning For Information Platforms

FORRESTER®

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
+1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com

MLDCs Scale Out Metadata And Policies For Modern Data Activation

Machine learning data catalogs (MLDCs) are more than a metadata management tool and marketplace. Standalone tools provide an enterprise hub across the ecosystem and solution- and platform-based catalog and metadata repositories. This hub combines a traditional data management business glossary, data stewardship, data preparation, and data marketplaces for a central platform to serve contextualized data. Machine learning (ML) is the glue that makes this happen.[1] ML automates the mundane aspects of understanding the data and applying policies, business rules, tags, and classifications. It provides introspection and inferencing to identify and anticipate error and conflict impacts. And it speeds up collaboration, data curation, and remediation with embedded intelligence and behavioral learning from a social media user experience (UX).

As a result of these trends, MLDC customers should look for providers that:

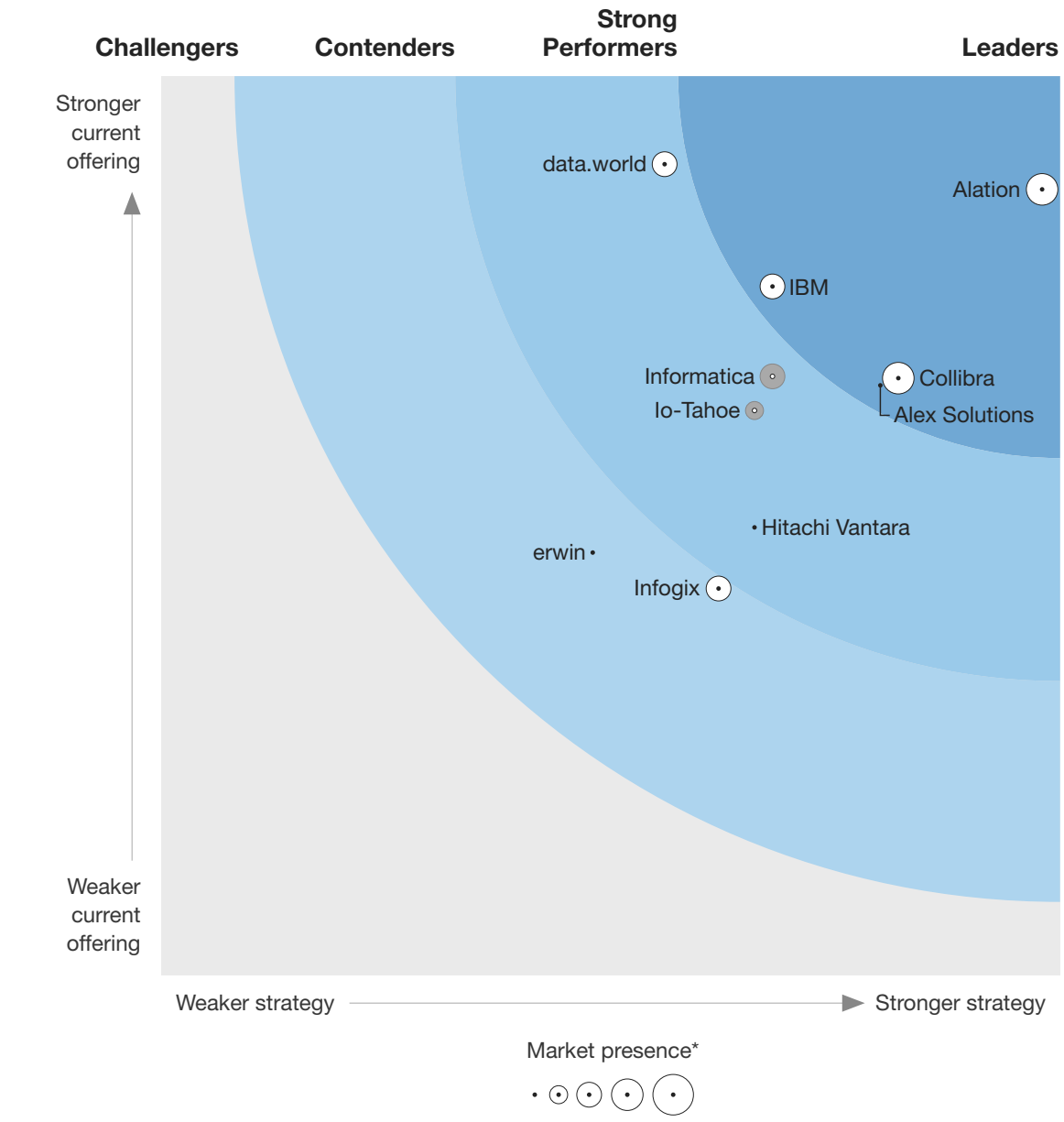
- **Power DataOps, data stewardship, and analytics process collaboration.**[2] UX, workflow, data intelligence, and embedded ML automates role-based data responsibilities and tasks while enabling communication and alerts. As data engineering evolves into a shared responsibility, integration with DevOps, collaboration, business process management (BPM), and solution platforms further reinforces the ability of data pros in all roles to work together.
- **Scale data intelligence and lineage across from metadata to the endpoint.** Data engineers will find better support to get beyond data understanding and support impact analysis, quality alerts, and root cause analysis. ML and metadata capture extend lineage to see complete data flows, including an understanding of endpoints such as data source, view, web, mobile device, and machine. In some tools, ML may inference the gaps in lineage.
- **Allow data and analytic professionals to work with all data.** MLDCs are advancing to support the variety of data that enterprises want to take advantage of. Connectors are becoming available for various content and media platforms. In 2018, these tools focused on structured data. Today, connectors and APIs are available to grab metadata from content and media. Some tools use natural language processing for rich metadata and context capture for policies and data protection.

Evaluation Summary

The Forrester Wave™ evaluation highlights Leaders, Strong Performers, Contenders, and Challengers. It's an assessment of the top vendors in the market and does not represent the entire vendor landscape. You'll find more information about this market in our reports on data management.

We intend this evaluation to be a starting point only and encourage clients to view product evaluations and adapt criteria weightings using the Excel-based vendor comparison tool (see Figure 1 and see Figure 2). Click the link at the beginning of this report on Forrester.com to download the tool.

FIGURE 1 Forrester Wave™: Machine Learning Data Catalogs, Q4 2020



*A gray bubble indicates a nonparticipating vendor.

FIGURE 2 Forrester Wave™: Machine Learning Data Catalogs Scorecard, Q4 2020

	Forrester's weighting	Alation	Alex Solutions	Collibra	data.world	erwin	Hitachi Vantara	IBM	Infogix	Informatica*	Io-Tahoe*
Current offering	50%	4.37	3.28	3.32	4.51	2.35	2.49	3.83	2.15	3.33	3.14
User experience	5%	4.00	2.70	3.30	5.00	2.70	1.90	3.30	2.00	3.80	4.00
Profiling and analytics	10%	3.80	3.80	3.00	4.20	2.60	2.60	3.40	2.20	3.40	3.80
Cataloging	10%	4.33	5.00	3.00	4.33	3.00	3.00	4.33	2.33	3.67	3.67
Stewardship and governance	5%	4.00	3.00	4.00	4.00	3.00	1.00	4.00	4.00	2.00	3.00
Collaboration	5%	5.00	3.00	3.00	5.00	3.00	3.00	3.00	3.00	3.00	1.00
Data activation and services	10%	3.50	3.00	2.50	4.50	2.50	2.50	4.00	3.00	3.00	2.50
Machine learning (ML)	20%	5.00	4.00	3.50	4.00	1.00	3.00	4.00	1.00	3.00	4.00
Data engineering	10%	3.50	1.50	2.00	5.00	4.50	1.50	3.50	1.00	3.50	2.50
Systems integration	20%	4.80	2.80	4.00	4.80	1.50	2.40	4.20	3.00	3.40	2.70
Security	5%	5.00	3.00	5.00	5.00	3.00	3.00	3.00	1.00	5.00	3.00
Strategy	50%	4.90	4.00	4.10	2.80	2.40	3.30	3.40	3.10	3.40	3.30
Product vision	30%	5.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	5.00
Delivery model	10%	5.00	5.00	5.00	1.00	5.00	5.00	5.00	3.00	5.00	3.00
Planned enhancements	40%	5.00	5.00	5.00	3.00	1.00	3.00	3.00	3.00	3.00	3.00
Market approach	10%	5.00	3.00	3.00	3.00	3.00	5.00	5.00	3.00	3.00	1.00
Performance	5%	5.00	3.00	3.00	1.00	3.00	3.00	3.00	3.00	5.00	3.00
Partner ecosystem	5%	3.00	3.00	5.00	5.00	3.00	1.00	3.00	5.00	5.00	1.00
Market presence	0%	4.00	1.00	4.00	3.00	1.00	1.00	3.00	3.00	3.00	2.00
Growth	50%	3.00	1.00	3.00	5.00	1.00	1.00	1.00	1.00	1.00	3.00
Revenue and customers	50%	5.00	1.00	5.00	1.00	1.00	1.00	5.00	5.00	5.00	1.00

All scores are based on a scale of 0 (weak) to 5 (strong).

*Indicates a nonparticipating vendor

Vendor Offerings

Forrester included 10 vendors in this assessment: Alation, Alex Solutions, Collibra, data.world, erwin, Hitachi Vantara, IBM, Infogix, Informatica, and Io-Tahoe (see Figure 3).

FIGURE 3 Evaluated Vendors And Product Information

Vendor	Product evaluated	Product version evaluated
Alation	Alation Data Catalog Platform	
Alex Solutions	Alex	4.0.0
Collibra	Collibra Data Intelligence Cloud	
data.world	data.world Enterprise Catalog	
erwin	erwin Data Intelligence Suite	10.1
Hitachi Vantara	Lumada Data Catalog	2019.3
IBM	IBM Watson Knowledge Catalog	3.0
Infogix	Data360	2020.7
Informatica	Informatica Enterprise Data Catalog	
Io-Tahoe	Io-Tahoe Data Catalog	

Vendor Profiles

Our analysis uncovered the following strengths and weaknesses of individual vendors.

Leaders

› **Alation provides an AI data expert for hire.** Alation exploits machine learning at every opportunity to improve data management, governance, and consumption by analytic citizens. Every data catalog function is underpinned by intelligence that learns from data patterns, queries, and data professional search and interaction. Customer references specifically call out the unique functionality of the Behavioral Analysis Engine, which assists data sourcing and preparation by guiding SQL scripting. Collaboration capabilities are another key reference customer delight. Alation is investing in them across the data catalog, including a complete overhaul of the UX.

Improvements in anomaly detection is a place for Alation to expand impact analysis and guided resolution. Customer references indicate a desire for end-to-end diagramming as well as better visibility into APIs, BPM, and workflows, along with enhanced connections between data lakes and Spark. Alation has several capabilities that support data and metadata syndication, sharing, and delivery, but this too is a place for growth. Alation has plans to broaden its support of modern data and analytic platforms, from expansion of data sources to richer integration with data flows, which will offset these points. For organizations hitting the accelerator on being an insights-driven business, Alation has demonstrated to its customers that it's the solid choice to intelligently harmonize data management, governance, and analytics.

- **Collibra expands data intelligence to connect data governance to business value.** Already exceling with Data Governance, Collibra extends its platform with deeper data intelligence. A recently announced partnership with Google Cloud Platform, a comprehensive privacy product, and advances in machine learning bring data from context to consumption. Reference customers describe having great communication with both customer support and product teams and receptiveness to product input for innovation. Customers specifically called out the data intelligence combined with graph technology. Additionally, they indicated that Collibra understands what businesses are trying to do with data, not just the IT side.

  Collibra's data catalog is comprehensive, but customer references find the user interfaces to be counterintuitive at times. Views of data models and data lineage are detailed but could be improved to better navigate and explore for anomalies, impact, and errors. Support for data science helps data scientists find data and understand it, but enablement can be lacking in workflows to manage models or integrate with data science tools. While regulatory privacy modules are available, Collibra does not provide regulatory content and instead provides a framework that allows customers and trained systems integrators to develop and operationalize their own content within the module. Organizations with complex and broad data governance, data literacy, and business intelligence (BI) and analytic support regularly choose Collibra for data democratization and business growth strategies.

- **Alex Solutions shows that data excellence can come from Down Under.** Alex Solutions brings business context to data for the data steward and data engineer. Its most notable capability is lineage profiling, which provides data flow, cross-system flow, application flow, and mobile flow to see data in the digital ecosystem. Its metadata management foundation, along with profiling and anomaly detection, ensures that data is trusted and not prone to data errors or consumption errors. Reference customers specifically call out the solution's flexibility and its ability to broadly scale and connect with a variety of data sources and data flows. They appreciate this scale to handle a variety of use cases.

  While foundational analytics and BI use cases are supported, data science is less so, according to customer references. This was evident in its fewer capabilities to integrate the data catalog with data science environments and notebooks as well as limited alignment between data scientists and DataOps. Customers give Alex Solutions high marks for its capabilities but would like more

support to train business users. Alex Solutions is moving from a regional footprint and expanding in North America to be more global while taking on more advanced use cases. Its machine learning is solid and its data catalog foundation impressive, making it a vendor to consider for organizations advancing their insights-driven business maturity.

› **IBM Watson Knowledge Catalog leans in on data activation and governance.** IBM Watson Knowledge Catalog is the inferencing engine to understand, classify, and govern data. IBM takes advantage of its cognitive IP and experience to bring reliable semantic intelligence, quality indicators, and collaborative experiences for all data and analytic professionals. Of particular note, Watson Knowledge Catalog is one of the few tools able to inference on documents, not just structured data, for classification and privacy. This sets the stage for other MLDCs to build the convergence bridge to unify the structured data and content worlds.

While the environment can be sold on its own, customers tend to invest in Cloud Pak for Data to power AI or migrate from existing IBM private cloud platforms. Customer references recognize that IBM Watson Knowledge Catalog is best for large, complex, and messy enterprise ecosystems. Challenges of metadata synchronization or migration tend to come from that environmental fact. Customer references would like a smoother implementation and transition, but they generally feel that IBM is highly responsive and does a good job to rectify these challenges during implementation. IBM Watson Knowledge Catalog is designed and deployed to handle enterprise-grade data to support scale-out AI strategies. Organizations that are shifting from advanced analytics to AI will find IBM Watson Knowledge Catalog a trusted capability to navigate, gather, prepare, and trust all information.

Strong Performers

› **Data.world emerges with a new vision for data management and governance.** A relative newcomer to the market, data.world brings a strong backbone to understanding, trusting, and activating data. It has moved quickly to step out of its niche data engineering position and serve data governance, analytic, and external marketplace objectives. The UX is intuitive, clean, and connected across key data roles (engineer, steward, analyst, and consumer). Combining its own machine learning with MANTA's lineage intelligence gives data.world an enterprise-grade profile and impact analytic capability appreciated by data engineers while supporting semantic representations for stewards and analysts.

Growing pains exist as data.world expands its features and functionality. Customer references indicated that a more unified experience between users could improve collaboration. References also mentioned other areas for improvement, such as better user education and advancements in visualizing performance. However, customers appear to be willing to work with data.world on these and appreciate the vendor's roadmap. Organizations seeking to start with a solid modern data platform foundation and grow to accommodate governance and analytic use cases will have a good partner in data.world.

- **Informatica EDC is the workhorse in the data fabric.** Informatica EDC is a proven data catalog for enterprise management of metadata. Strong partnerships with Microsoft and Amazon Web Services have helped EDC become a leading data catalog for the modern data warehouse and data fabric. A vast number of connectors to both long-standing and new data and data science platforms gives EDC a sound foundation to discover, profile, and inference a rich set of metadata, quality status, and lineage intelligence. EDC sits on a Hadoop Spark foundation and leverages both Hive and HBase to build the relational database of metadata captured. Additionally, it supports data intelligence with time-series capabilities for anomaly, change, and impact tracking. Graph technology helps manage the data source relationships. Machine learning does the heavy lifting for discovery, classification, and lineage analysis.

  Customers choose EDC specifically to support the data engineering and integration foundation for migration and standing up their data lakes and data fabrics across hybrid cloud and on-premises architectures. However, customers say they need more data stewardship support and generally acquire another data catalog to fill this data governance gap. Additionally, while EDC sits on a modern scalable architecture, customers cite concerns about performance and cost to support operational and event-driven use cases as barriers for their digital strategies dependent on IoT and service mesh architecture.[3] Forrester recommends Informatica EDC for enterprises looking to move to the cloud to advance their insight agendas and supply and holistically manage data for analytics and data science at scale. Informatica declined to participate in the full Forrester Wave evaluation process.

- **Io-Tahoe for automated data intelligence.** The Io-Tahoe data catalog was the first product release of the vendor's automated data platform. The strategy was to have a foundation of metadata management and services that would drive data automation across an enterprise environment. Semantics, lineage, anomaly detection, data statistics, and metadata capture are all central to the data catalog. Built around deep machine learning intelligence, the data catalog can understand the data, data flows, and source systems. It provides the detail needed by DataOps with the understandability for data stewards and data consumers. Customers rely on the profiling capabilities to support both data migration and modernization. But where Io-Tahoe really demonstrates value is the insight to interpret the data's business impact for quality, regulatory, and privacy governance.

  Io-Tahoe is enhancing its data catalog to support more automated data quality and data pipeline building. As such, the data catalog is evolving to be part of a broad platform for data. The benefit will be the ability to gain capabilities within Io-Tahoe's own ecosystem even as it has an API framework to span cloud and on-premises systems. Today, Io-Tahoe is still a growing company and transitioning from a product to a platform value proposition. A strong vision and roadmap are evident. However, customers should be prepared to grow with Io-Tahoe. For enterprises looking to leap ahead with machine learning for automation at the core of data management, Io-Tahoe Data Catalog is a logical choice. Io-Tahoe declined to participate in the full Forrester Wave evaluation process.

› **Hitachi Lumada Data Catalog is building the bedrock for the data fabric.** The recent acquisition of Waterline Data filled the gap in Hitachi Vantara's data fabric platform to understand, govern, and activate data. One of the first data catalogs to use machine learning, Lumada continues to invest in ML and incorporate more behavioral intelligence to drive automation and data prescriptions. Lineage analysis is the most differentiating feature, allowing data professionals and analysts to infer hidden lineage and transformations for a complete view of data flows and ownership. Customer references also mention that Lumada is simple to use, and its business users provide positive feedback.

Lumada is still evolving its capabilities to align with the way data governance and DataOps teams work. Today, data engineering is still primarily supported through the Pentaho environment, but it has added connectors to Jira to improve requests and task management. These connectors will also support stewardship and custodianship remediation. Customer references would like to see improvements in collaboration and better glossary capabilities to see descriptions, annotations, and metadata together. Hitachi Lumada Data Catalog is a strong candidate to help analysts and data consumers find and understand their data. Organizations seeking to first enable data democratization and build back to the data fabric can achieve these objectives with Lumada.

Contenders

› **Infogix gives data stewards a platform to build data value.** Infogix puts data leaders at the center of their data. Ultimately, Infogix's data catalog is meant to deliver data intelligence and quantify the value data provides to the business. Data governance is a core driver of functionality and oriented from business metrics down to the data. This translates into a clean UX for data stewards and consumers as well as personalized data intelligence dashboards. With data economics in mind, Infogix has begun to add data science and engineering support with Python generation. Customer references reported being pleased with the data catalog tool and cite how easy it is to work with Infogix.

Improvements for technical and analytic support are needed. The data catalog has most of the functionality checked off in terms of machine learning, integration, search, and profiling. However, these capabilities are still emerging as Infogix commits to further development into 2021. Customer references indicate that in general the tool has lived up to expectations but has a limited ability to deal with diverse ecosystems and handle large volumes of data. Organizations will find Infogix a good partner to embark on an insights-driven business journey by driving data trust, literacy, and proof of value.

› **Erwin demonstrates how catalogs connect data engineers and stewards.** When data teams hear the name erwin, they first think of data modeling. Over the last three years, however, a combination of development and acquisition has given erwin a data engineering bench grounded in its MLDC. Robust support for workflow, data search, metadata capture, and data operations brings

data professionals up to speed quickly with their data. Customer references called out automation and the ability to share business metrics (to drive data literacy) as key reasons to implement erwin. They also called out erwin's strong data mapping and lineage.

The vendor's ML algorithms are less sophisticated, although they cover the key areas of usage, semantics, and behavioral intelligence and support. Where machine learning has helped other vendors speed up time-to-value, erwin customers indicate a six- to eight-month time frame. Additionally, customer references think there is room for erwin to develop its connectors and work out kinks. However, references were overall very satisfied and willing to work with erwin as the company grows. For existing and new customers alike, erwin is worth taking another look at for organizations prioritizing data modernization, cloud, and aligning IT with business data objectives and governance.

Evaluation Overview

We evaluated vendors against 39 criteria, which we grouped into three high-level categories:

- **Current offering.** Each vendor's position on the vertical axis of the Forrester Wave graphic indicates the strength of its current offering. The key criteria for these solutions are user experience, profiling and analytics, cataloging, stewardship and governance, collaboration, data activation and services, machine learning, data engineering, systems integration, and security.
- **Strategy.** Placement on the horizontal axis indicates the strength of the vendors' strategies. We evaluated product vision, delivery model, planned enhancements, market approach, performance, and partner ecosystem.
- **Market presence.** Represented by the size of the markers on the graphic, our market presence scores reflect each vendor's growth and revenue.

Vendor Inclusion Criteria

Forrester included 10 vendors in the assessment: Alation, Alex Solutions, Collibra, data.world, erwin, Hitachi Vantara, IBM, Infogix, Informatica, and Io-Tahoe. Each of these vendors has:

- A standalone MLDC offering.
- More than $1 million in revenue.
- Presence in three major geographies (North America, Europe, and Asia Pacific).
- A complete set of machine learning capabilities.

Supplemental Material

Online Resource

We publish all our Forrester Wave scores and weightings in an Excel file that provides detailed product evaluations and customizable rankings; download this tool by clicking the link at the beginning of this report on Forrester.com. We intend these scores and default weightings to serve only as a starting point and encourage readers to adapt the weightings to fit their individual needs.

The Forrester Wave Methodology

A Forrester Wave is a guide for buyers considering their purchasing options in a technology marketplace. To offer an equitable process for all participants, Forrester follows The Forrester Wave™ Methodology Guide to evaluate participating vendors.

In our review, we conduct primary research to develop a list of vendors to consider for the evaluation. From that initial pool of vendors, we narrow our final list based on the inclusion criteria. We then gather details of product and strategy through a detailed questionnaire, demos/briefings, and customer reference surveys/interviews. We use those inputs, along with the analyst's experience and expertise in the marketplace, to score vendors, using a relative rating system that compares each vendor against the others in the evaluation.

We include the Forrester Wave publishing date (quarter and year) clearly in the title of each Forrester Wave report. We evaluated the vendors participating in this Forrester Wave using materials they provided to us by July 17, 2020, and did not allow additional information after that point. We encourage readers to evaluate how the market and vendor offerings change over time.

In accordance with The Forrester Wave™ Vendor Review Policy, Forrester asks vendors to review our findings prior to publishing to check for accuracy. Vendors marked as nonparticipating vendors in the Forrester Wave graphic met our defined inclusion criteria but declined to participate in or contributed only partially to the evaluation. We score these vendors in accordance with The Forrester Wave™ And The Forrester New Wave™ Nonparticipating And Incomplete Participation Vendor Policy and publish their positioning along with those of the participating vendors.

Integrity Policy

We conduct all our research, including Forrester Wave evaluations, in accordance with the Integrity Policy posted on our website.

Endnotes

[1] See the Forrester report "Reclaim Your Semantic Desert To Monetize Your Data."

[2] DataOps: data operations.

[3] IoT: internet of things.

We work with business and technology leaders to drive customer-obsessed vision, strategy, and execution that accelerate growth.

PRODUCTS AND SERVICES

- Research and tools
- Analyst engagement
- Data and analytics
- Peer collaboration
- Consulting
- Events
- Certification programs

Forrester's research and insights are tailored to your role and critical business initiatives.

ROLES WE SERVE

Marketing & Strategy Professionals

CMO
B2B Marketing
B2C Marketing
Customer Experience
Customer Insights
eBusiness & Channel Strategy

Technology Management Professionals

CIO
Application Development & Delivery
- Enterprise Architecture

Infrastructure & Operations
Security & Risk
Sourcing & Vendor Management

Technology Industry Professionals

Analyst Relations

CLIENT SUPPORT

For information on hard-copy or electronic reprints, please contact Client Support at +1 866-367-7378, +1 617-613-5730, or clientsupport@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.