Market Guide for Data Preparation

Published: 14 December 2017 **ID:** G00315888
Analyst(s):
Ehtisham Zaidi, Rita L. Sallam, Shubhangi Vashisth

Summary

Data preparation — the most time-consuming task in analytics and BI — is evolving from a self-service activity to an enterprise imperative. We profile 28 data preparation tools for data and analytics leaders to consider to accelerate agile data preparation for a range of distributed content authors.

Overview

Key Findings

- The market for data preparation has now evolved from tools supporting only self-service use cases into platforms that enable data and analytics teams to build agile and searchable datasets at an enterprise scale for distributed content authors.
- Most vendor offerings support data profiling, data exploration, transformation, modeling and curation, and metadata support. More than 80% of the vendors surveyed embed some data cataloging features and offer varying degrees of machine-learning capabilities.
- The market is crowded with a range of choices, from stand-alone specialists to vendors that embed data preparation as a capability into analytics and BI, data science, or enterprise data integration platforms. Although accelerating the shift toward broadly deployed modern analytics and data science, these tools, if unmanaged, can introduce multiple versions of the truth.

Recommendations

Data and analytics leaders modernizing their data management and analytics strategies:

- Develop a deployment strategy for data preparation to enhance user understanding of data, reduce data preparation efforts and increase agility. Evaluate vendors based on capabilities, integration points, pricing and roadmaps.
- Create a formal process for vetting and reusing models developed by business users, for operationalizing data preparation flows, and for incorporating them into the enterprise data integration workflow, as warranted. Recognize that, while data preparation tools can be used for an increasing number of new data integration use cases, they do not yet replace the need for enterprise data integration solutions for all requirements.

- Investigate your data preparation vendors' roadmap on their current or planned support for extended data preparation capabilities to improve the interactive experience, facilitate timely insights and enhance enterprise readiness. Examples include the inclusion of data science libraries, more-intuitive data preparation workflows, improved governance, collaboration, machine learning and cataloging.

Strategic Planning Assumptions

By 2020, data preparation tools will be used in more than 50% of new data integration efforts for analytics.

By 2023, machine-learning-augmented master data management (MDM), data quality, data preparation and data catalogs will converge into a single modern enterprise information management (EIM) platform used for the majority of new analytics projects.

By 2019, data and analytics organizations that provide agile, curated internal and external datasets for a range of content authors will realize twice the business benefits as those that do not.

Market Definition

This document was revised on 10 January 2018. The document you are viewing is the corrected version. For more information, see the Corrections page on gartner.com.

Rapid Proliferation of Data and Analytics Demand Creates the Need for Data Preparation

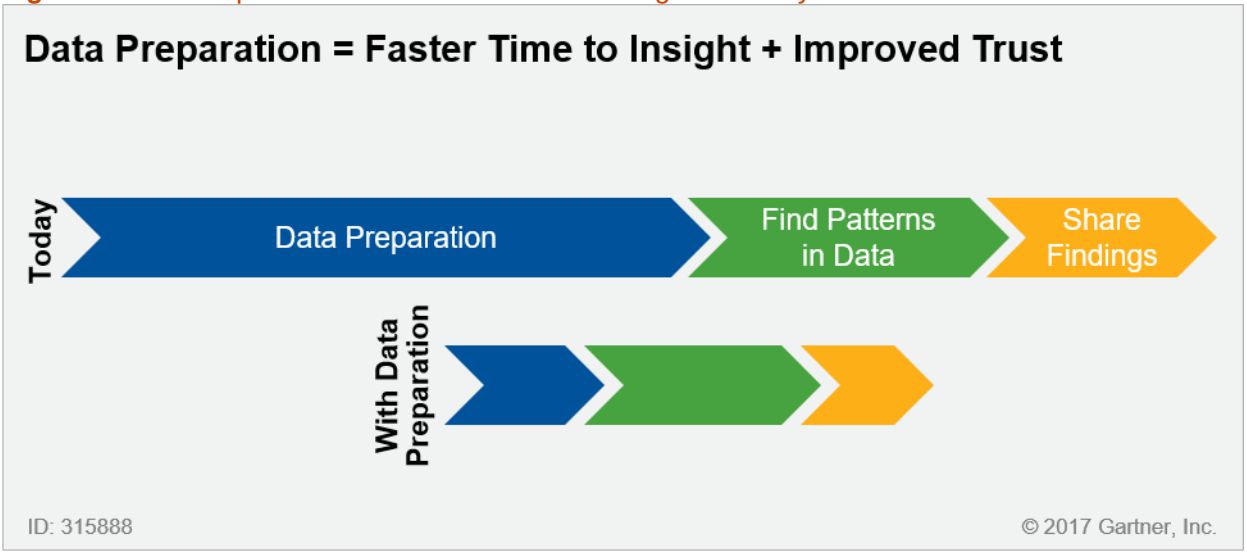
The business is demanding faster time to insight than ever before to remain competitive, particularly as more industries are facing digital disruption. As a result, analytics is becoming more pervasive across the enterprise and those insights are being derived from larger numbers of diverse data sources, both internal and external to the enterprise, with varying degrees of trustworthiness. Amid this increasing complexity and business urgency, business people are challenged to get the insights they need in time without IT assistance. Data and analytics leaders are struggling to respond to this urgent need due to their over-reliance on IT-centric tools for finding, cataloging and transforming relevant data and making it accessible to the growing number of distributed users in the enterprise both inside and outside of centralized data and analytics teams. Due to this, organizations report that they spend more than 60% of their time in data preparation, leaving little time for actual analysis. And, in a recent Gartner Research Circle study, respondents indicated that they are most likely to automate data integration (60%) and data preparation (54%) in the next 12 to 24 months (see the Evidence section).

The Data Preparation Process

In response to this problem, data preparation tools from modern analytics and BI, data science, and even data integration tool providers have emerged to allow users of varying skills levels to access, integrate and transform data for their own analysis. They rapidly speed up time to insight by allowing users to reduce the complexity of data preparation and transformation, find patterns

and anomalies in their integrated datasets, and share their findings for further analysis without extensive IT support or significant coding knowledge (see Figure 1). However, these tools have varying degrees of support (e.g., data lineage, impact analysis and metadata management) for informing the work of information governance and stewardship when building large and complex data models.

Figure 1. Data Preparation Reduces the Time to Insight for Analytics



Source: Gartner (December 2017)

Without processes in place, this can rapidly lead to a potential governance problem and multiple versions of the truth where multiple teams with a range of users (with varying skills levels in data integration, modeling and analytics) could end up using separate point solutions to prepare and harmonize datasets for analytics without a coordinated data governance program. This could lead to unmanageable complexity when dealing with different types of data in the context of truth-based and trusted-based policies (see "Reset Your Information Governance Approach by Moving From Truth to Trust").

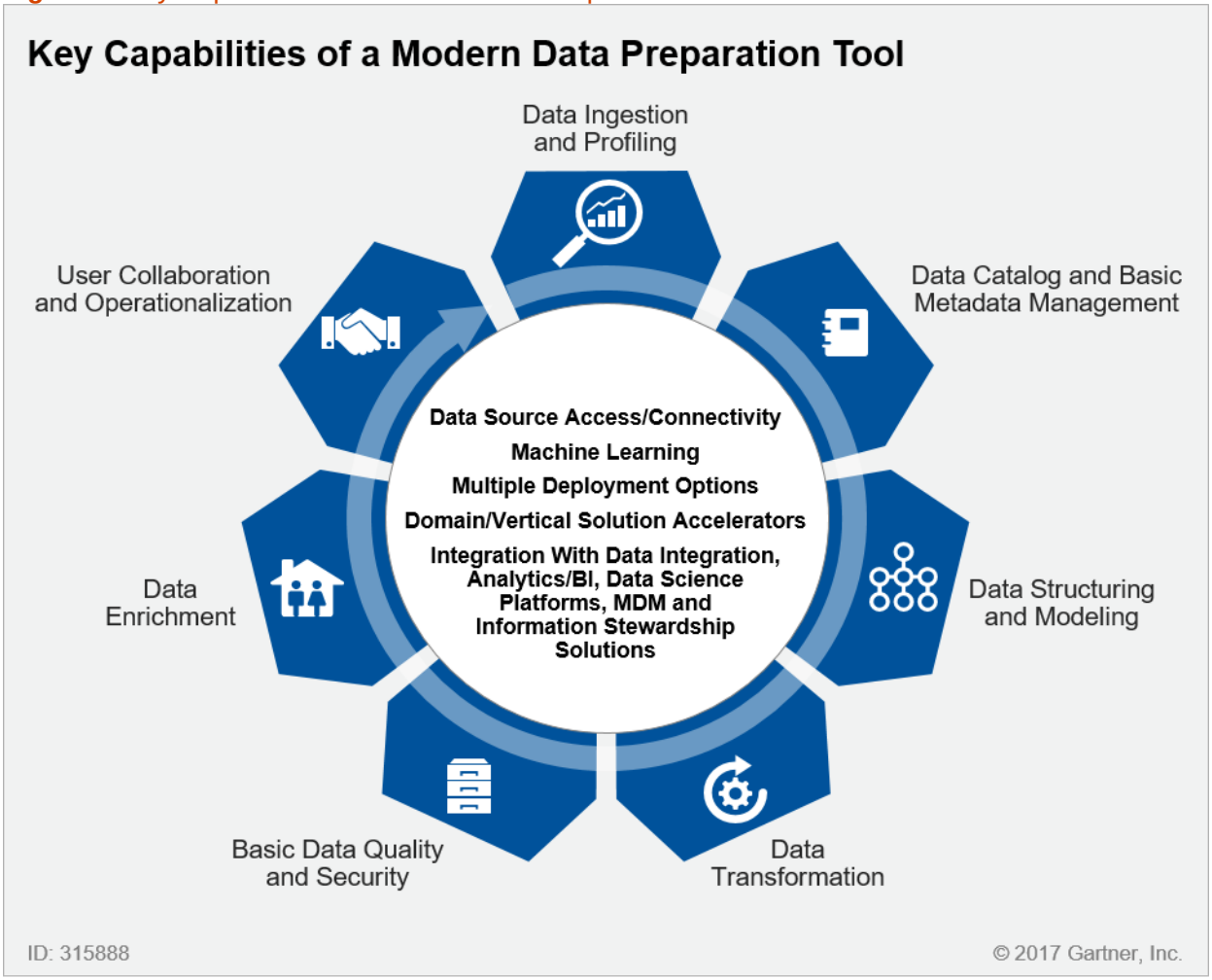
Data preparation is an iterative-agile process for exploring, combining, cleaning and transforming raw data into curated datasets for self-service data integration, data science, data discovery, and BI/analytics.

To perform data preparation, data preparation tools are used by analysts, citizen data scientists and data scientists for self-service. The tools are also used by citizen integrators and data engineers for data enablement to reduce the time and complexity of interactively accessing, cataloging, harmonizing, transforming and modeling data for analytics in an agile manner with metadata and lineage support. These tools can provide data access for use in mostly analytical tasks that include storage, logical and physical data modeling, and data manipulation for data visualization, data integration and analytics. Some tools support machine-learning algorithms that can recommend or even automate actions to augment and accelerate data preparation.

Capabilities Needed in a Modern Data Preparation Tool

In order to support the ever-evolving use-case requirements of an organization, modern data preparation tools must support and enable the high-level, core data preparation capabilities listed below (see Figure 2).

Figure 2. Key Capabilities of a Modern Data Preparation Tool



Source: Gartner (December 2017)

1. **Data ingestion, exploration and profiling:** A visual environment that enables users to interactively prepare, search, sample, profile, catalog and inventory data assets, as well as tag and annotate data for future exploration. Advanced features include autoinference, discovering and suggesting sensitive attributes, identifying commonly used attributes (for example, geodata and product ID), doing semantic reconciliation, discovering and recording data lineage of transformations, and autorecommending sources to enrich the data.
2. **Data cataloging and basic metadata management:** Supports creating and searching metadata, cataloging of data sources, transformations, user activity against the data source, data source attributes, data lineage and relationships, and APIs to enable access to the metadata catalog for auditing or other uses. Through the use of analytics on the raw data, the models are derived and generated bottom up instead of designed top down. It is a continuous process of accumulating metadata based on the actual use of data — it is a living construct. Gartner is seeing some data preparation tools incorporate a data catalog within

the data preparation workflow. These focused data catalogs are point solutions that allow users to inventory the data assets integrated by these solutions. They must not be confused with overall enterprise metadata management solutions that have a more broad scope of managing metadata across the overall data and analytics program, and not just limited to the data preparation needs (see "Magic Quadrant for Metadata Management Solutions").

3. **Data structuring, modeling and transformation:** Supports data mashup and blending; data cleansing; filtering; and user-defined calculations, groups and hierarchies. This includes agile data modeling/structuring that allows users to specify data types and relationships. More-advanced capabilities automatically deduce or infer the structure from the data source, and generate semantic models and ontologies, such as logical data models and hive schemas.
4. **Basic data quality and security:** Integration with tools supporting information governance and stewardship and capabilities for data encryption, user permissions and data lineage. This also includes security features, such as data masking, platform authentication and security filtering at the user/group/role level, as well as through integration with corporate LDAP and/or Active Directory systems, SSO, source system security inheritance, row- and column-level security, and logging and monitoring of data usage and assets. Some tools offer visualizations within the data preparation flow to show data distribution, outliers and other relationships.
5. **Data enrichment:** Support for basic data enrichment capabilities including entity extraction (capturing of attributes from the data integrated using the data preparation tool); attribute development (allowing process experts to develop the attribute set for integrated data based on requirements of their industry or domain); and improving the data enrichment cycle through machine learning for future use as more datasets get added to enhance productivity of analysts and other users.
6. **User collaboration:** Facilitates the sharing of queries and datasets, including publishing, sharing and promoting models with governance features, such as dataset user ratings or official watermarking. This is where it is important for self-service data preparation tools to not exist in isolation. They need to have the ability to catalog, share and govern metadata either through their own embedded data cataloging features or through their ability to share metadata with the more complete and broad metadata management tools. These incorporate other critical capabilities other than just data cataloging, such as data lineage, business glossary, rules management, impact analysis and semantic frameworks that allow the data preparation flow to be governed, managed and audited by IT for data quality and governance before promoting them.
7. **Data source access and connectivity:** APIs and standards-based connectivity, including native access to cloud application and data sources, enterprise on-premises data sources, relational and unstructured data, NoSQL, Hadoop, and various file formats (XML, JSON, .csv), as well as native access to open, premium or curated data. Modern data preparation tool vendors are also adding support for streaming data such as machine and Internet of Things (IoT) data to their portfolio.
8. **Machine learning:** Use of machine learning and artificial intelligence (AI) to improve and, in some cases, even automate the data preparation process. Some tools provide algorithms

to enable users to identify data structures, schemas and relationships at various levels of granularity, and the ability to structure the datasets upon initial data ingestion. These algorithms develop over time to recommend with more precision the most accurate data sources, joins, transformations and enrichments, making the user more productive. Machine learning also allows users to understand when they can operationalize their data preparation flows and promote them to departmentwide/organizationwide data integration flows through the ability to share metadata in an incremental-bidirectional manner with external metadata management, data quality and governance tools.

9. **Deployment models:** These tools can be deployed either in the cloud, on-premises, or across both cloud and on-premises. This latter hybrid approach allows users to leave data on-premises in place for processing, rather than moving it to the self-service data preparation platform, either in the cloud or on-premises. A small number of vendors also support multicloud deployments.

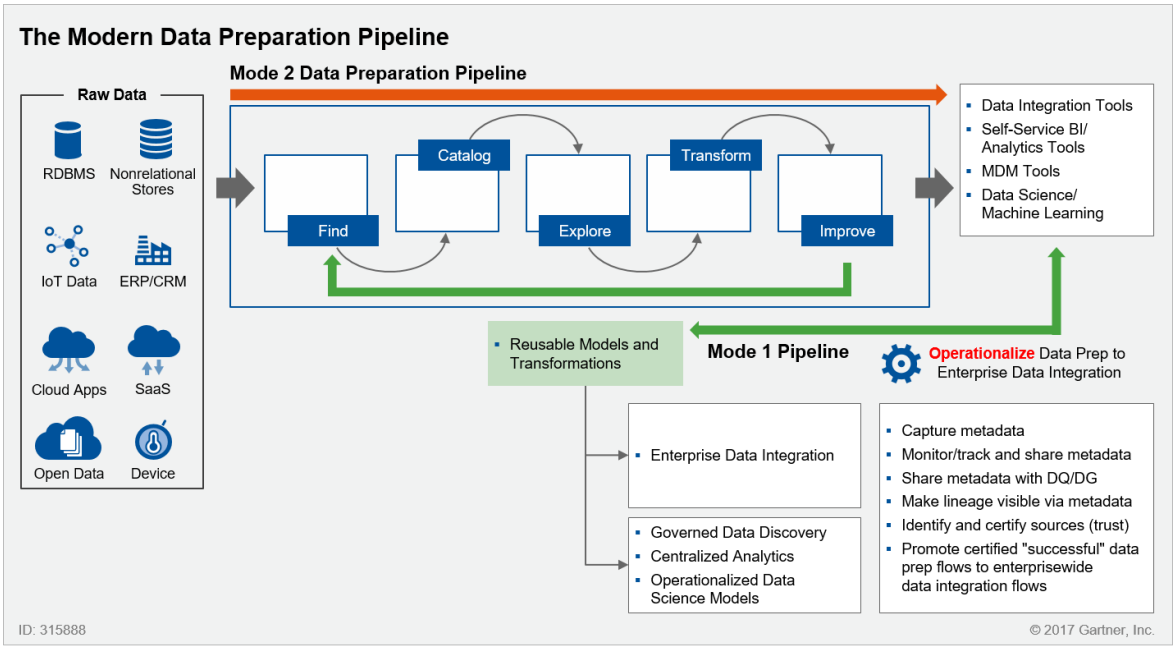
10. **Domain- or vertical-specific offerings or templates:** Packaged templates or offerings for domain- or vertical-specific data and models that can further accelerate time to data preparation and insight. This is particularly helpful for a number of difficult-to-use syndicated datasets.

11. **Integration with other data Integration, analytics/BI and data science platforms, MDM, and information stewardship solutions:** The ability to integrate harmonized datasets with data integration and analytics/BI and data science platforms through APIs, web services or native support for partner file formats (for example, .tde for Tableau, .qvd for Qlik and .pbi for Microsoft Power BI). At the same time, the information gleaned from the work of data discovery can help inform the work of information governance (which is itself not a technology but a business discipline and process) and share discovered metadata and data with an information stewardship solution that can help improve policy enforcement (see "Market Guide for Information Stewardship Applications").

Operationalization of Data Preparation Flows

Most data preparation tasks start as Mode 2 activities in the data preparation pipeline (see Figure 3) where business users (e.g., citizen data scientists and citizen data integrators, for example) connect to multiple, frequently changing data sources and where the time to insight is more critical as compared to upfront governance. These tasks originate from the business need to experiment with new data sources and data types where the business is uncertain about the eventual success and viability of the experiment and wants to quickly test hypothesis without having to wait for IT support. This is where the self-service capabilities of data preparation tools allow the business to break down the barriers of having to understand coding and data management for vastly reduced time to solution.

Figure 3. Operationalization of Data Preparation — The Modern Data Preparation Pipeline



Source: Gartner (December 2017)

Once the Mode 2 experiment is successful and the business wants to operationalize (promote to system of record) its findings, it must ensure that IT puts in place the required sandbox and promotion processes based on central governance strategies and rules. IT should ensure it has the capabilities in place to capture and share metadata from these data preparation flows bidirectionally. There should also be other data management tools in place like metadata management, data quality, data integration and data governance, to monitor and analyze the reuse of data and then recommend and even automate the promotion of these Mode 2 data preparation flows to broader reusable Mode 1 transformations via enterprise data integration when warranted. This process will ensure that the organization has the ability to promote initiatives that started as "self-service" to governed data discovery, centralized analytics, operationalized data science and other mission-critical initiatives when needed. It is important that data preparation tools support the "operationalization of data preparation flows" to ensure trust and governance and enterprise use cases.

Market Direction

There are a number of trends that will continue to drive adoption of data preparation tools:

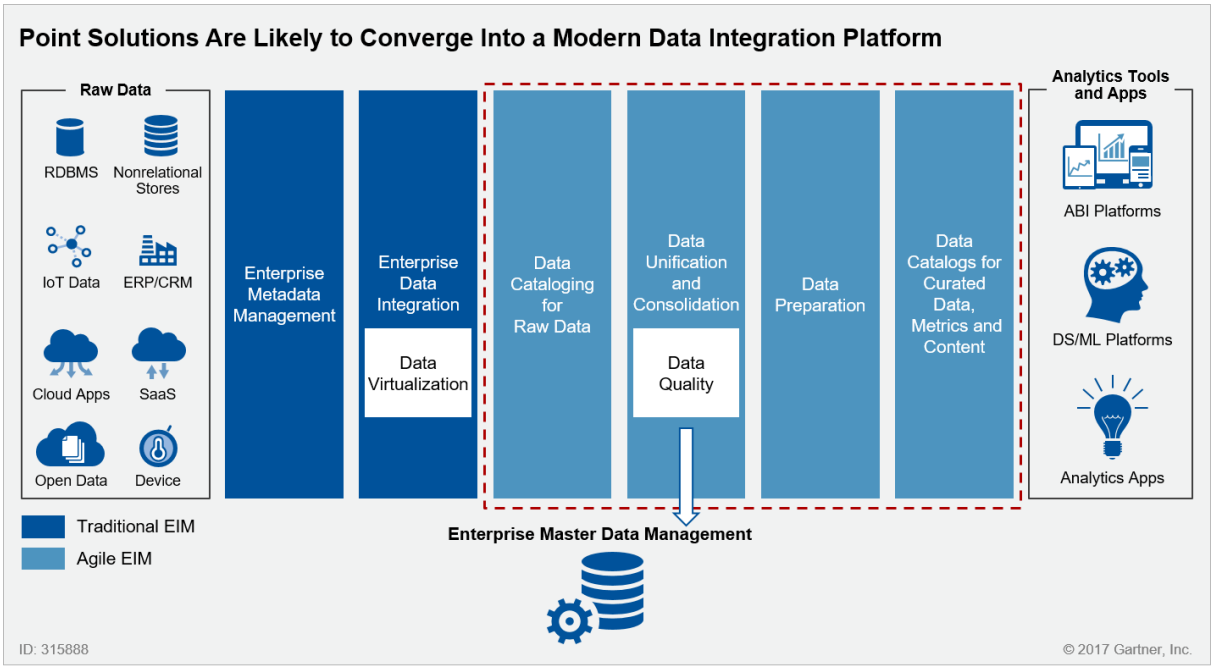
- **The evolution from self-service data preparation to enterprise data preparation:** The market for data preparation tools is evolving from self-service to enterprise in a similar way to the market for modern BI and analytics platforms. Initially, because of their agility and ease of use, data preparation tools started out being used only for self-service use cases by analysts and data scientists to accelerate the preparation of data for interactive analysis and data science. As self-service analytics was the primary use for these tools previously, Gartner initially called this market "self-service data preparation." Self-service data preparation capabilities are now available as a key embedded feature of modern BIA platforms, data science and machine-learning platforms and data integration tools. However, over time, specialist, stand-alone data preparation tools have emerged as the need

to create trusted, complex data models from increasingly larger numbers of datasets for many content authors using a range of different tools has grown beyond the capabilities of most embedded self-service data preparation tools.

- These stand-alone tools now enable data engineers and data analysts on centralized data and analytics teams to quickly prepare curated and governed datasets at scale for use across the enterprise. We are also seeing central data and analytics team using these tools for enterprise integration use cases or to delegate some responsibility for business-oriented integration tasks to less skilled IT or non-IT roles, freeing up time for data and analytics specialists to focus on challenging requirements elsewhere. Hence, there has been an evolution of data preparation tools that initially supported only self-service use cases to tools that can support centralized enterprise use cases as they have become a key enabling component of modern, distributed and trusted analytics.
- **Data preparation and machine learning:** As data becomes more complex, it is virtually impossible or time-prohibitive to manage, clean, harmonize and shape it manually. Machine learning has become a critical feature to operationalize and automate time-consuming and error-prone data preparation tasks (see "Augmented Analytics Is the Future of Data and Analytics" and "Rebalance Your Integration Effort With a Mix of Human and Artificial Intelligence"). In our survey, we see that most data preparation tool vendors (about 70% of the surveyed vendors) have already incorporated some machine-learning algorithms into their data preparation tools to improve and make the data preparation process more productive.
- **Convergence:** There is an increasing convergence of capabilities for data preparation with relationship discovery and entity resolution tools that can support upstream data unification and consolidation (for example, Tamr and Reltio), tools supporting data profiling and data quality capabilities (for example, Informatica and Talend), tools supporting stream data integration capabilities (for example, Striim and StreamSets), and tools supporting data cataloging capabilities (for example, Alation and Waterline Data). These activities are all an important part of the overall data preparation pipeline and are often preferred by buyers as integrated features, rather than as unintegrated or loosely integrated point solutions. Many vendors of each of these individual components are adding or acquiring features of the other to create a more comprehensive workflow (e.g., Alteryx, Unifi, Trifacta, Paxata, Datameer and Datawatch, among others).

Figure 4 shows how these capabilities are likely to converge and complement Mode 1, system-of-record enterprise metadata management, enterprise MDM and enterprise data integration tools (shown as "Traditional EIM" in the figure) into a modern enterprise data integration architecture (shown as "Agile EIM"). Of the 26 vendors surveyed for this research, over 80% claim to have a data catalog incorporated within their existing data preparation offering or enable the use of a catalog using a third-party metadata management tool using API access (see the Representative Vendors section for more analysis).

Figure 4. Enterprise Information Management Market Convergence



Source: Gartner (December 2017)

- **Cloud as a service, hybrid and multicloud:** Buyers are looking for true platform as a service (PaaS) data preparation tools that provide the flexibility and scalability of cloud data integration. Moreover, organizations need the flexibility to deploy data preparation and harmonization workloads where it makes the best sense, without having to move data first (to prepare and transform it). Support for all deployment modes, whether single/multicloud, on-premises or even hybrid, is becoming a critical enterprise requirement. We witness that a large majority (over 75%) of our surveyed data preparation tool vendors provide all three deployment options (i.e., on-premises, cloud and hybrid), which is in-line with the market demand.

- **The rise of the data engineer:** Data preparation tools have initially been used primarily by analysts and data scientists to expedite the preparation of data as part of iterative analysis. However, the role of data engineer, a centralized data specialist, is emerging in IT, and is using these tools to speed the creation of curated, trusted data for a range of distributed analytics content authors. As self-service and data science become more pervasive, data and analytics leaders have an opportunity to shift and expand their teams' role from being analytics and BI content creators to user and data enablers with this new data engineer role (see "Organizing Your Teams for Modern Data and Analytics Deployment").

- **The rise in use of data preparation tools in data lake use cases:** The business value of a data lake is entirely determined by the skills of the users using the lake. At its core, the data lake is simply a data storage strategy, not an analytical platform (see "Defining the Data Lake"). The data lakes that are typically used for data science use cases can benefit from data preparation tools to empower data engineers to transform and prepare data for further analysis (for downstream data science or analytics use cases) and to support their data scientists and citizen data scientist counterparts. These tools are rapidly assisting data engineers to break down the barriers to data lake adoption by helping them integrate and

transform data without having to master coding (SQL, Python, etc.; see "Three Architecture Styles for a Useful Data Lake").

- **Partnerships with data virtualization tools:** Data preparation tools are being increasingly deployed with and enabled through data virtualization technology, which allows them to connect to several different data sources and integrate data without having to physically move data into silos (see "Adopt Data Virtualization to Improve Agility and Bimodal Traits in Your Aging Data Integration" and "Market Guide for Data Virtualization"). Although most data preparation tools come with their own semantic virtual tier technology to integrate data sources, these are limited in their overall capabilities on performance optimization, dynamic query optimization, support for push-down processing to the underlying data sources, and their lack of persistence and write-back features to the underlying data stores. This is where data preparation tools are partnering with established data virtualization tools to deliver integrated data for further transformation and analysis.
- **Enterprise features for scale, security and governance:** Data preparation is no longer limited to self-service use cases. As data becomes more complex, from both internal and external sources, and analytics becomes more distributed and mission-critical, the need to secure, govern and scale accessibility to curated and trusted data — in an agile and timely way — become critical imperatives. Data preparation tools alone cannot fulfill these requirements, but as they get more mature, we anticipate increased use of them by new and emerging solutions related to information governance and stewardship through better overall integration and through more openness for metadata exchange.

Overall Market Outlook

The market for data preparation is currently estimated to be around $780 million in software revenue and will continue to grow and experience a healthy 18.5% CAGR (from 2016 through 2021), reaching an estimated value of $1.50 billion by 2021 (see "Forecast: Modern Business Intelligence Platforms by Select Functionality, Worldwide, 2016-2021").

Overall, we expect the market to continue consolidating over the next three to five years, where many of the stand-alone data preparation tools will either have expanded to provide enterprise-level, end-to-end analytical or data integration tool capabilities, or have been integrated with other data integration or analytics vendors that want to include comprehensive data preparation capabilities in their platform. However, there is also a growing IT/centralized data and analytics team use case for specialist data preparation tools that can provide agile and sanctioned data access in support of a range of complex data integration and analytics use cases for diverse data types, where there will continue to be room for differentiation and innovation. Data preparation tools can also continue to differentiate by supporting specialized data integration use cases (for example, IoT data preparation) or by improving the ease of data preparation (for example, through machine-learning algorithm integration), which continues to be a gap in the market (see "Hype Cycle for Data Management, 2017").

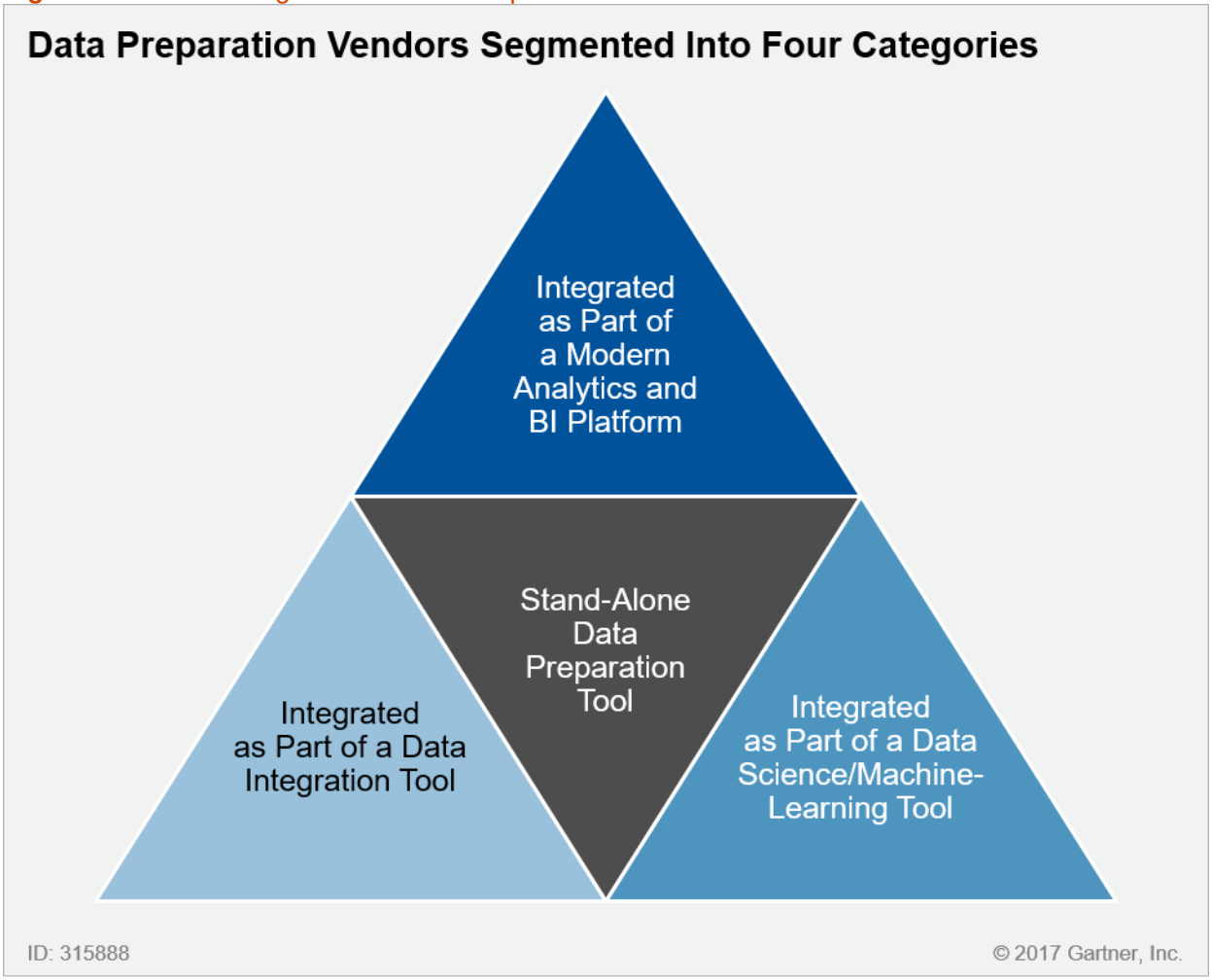
Market Analysis

The data preparation market is currently in a state of flux and is expanding with a range of vendor choices, from stand-alone specialists to vendors that embed self-service data preparation with analytics and BI, data science, and/or traditional data integration tools. It will likely consolidate over the next three to five years.

The Four Segments of the Data Preparation Tool Market

This Market Guide highlights data preparation vendors that are segmented into four categories (see Figure 5).

Figure 5. The Four Segments of Data Preparation



Source: Gartner (December 2017)

Category No. 1: Stand-alone data preparation tool. Vendors in this category sell data preparation as a stand-alone offering. Stand-alone vendor offerings focus on enabling tighter integration with downstream processes, such as API access and support for multiple analytics and BI and data integration tools. The stand-alone data preparation vendors are the focus of this Market Guide and have been analyzed in detail in the Representative Vendors section.

Data preparation is also embedded across the entire analytic workflow:

Category No. 2: Integrated as part of a data integration tool. Vendors here are focused on data integration and management that have added self-service data preparation to their product portfolios. This is done either by embedding data preparation capabilities to their existing data integration tools portfolio or as separate data preparation tools that can be purchased to support their data integration tools. They often offer some level of integration and promotability of data models between the data preparation and existing data integration tools. Data integration tool vendors such as Informatica, Lore IO, Hitachi Vantara (formerly Pentaho), SAS, Talend and TMMData have or are all adding data preparation to their enterprise data integration portfolios — those that have a stand-alone product are listed in the Representative Vendors section.

Category No. 3: Integrated as part of a modern analytics and BI platform. These integrated data preparation vendor offerings focus on data preparation capabilities as part of an end-to-end analytics process, with broader analytics and BI and content creation capabilities. All modern analytics and BI vendors have some embedded data preparation capabilities — this is a critical capability (see "Magic Quadrant for Business Intelligence and Analytics Platforms" and "Other Vendors to Consider for Modern BI and Analytics" for all vendors with this capability). Tableau (which will introduce a stand-alone data preparation product in 2018), Qlik, Microsoft Power BI, MicroStrategy, Oracle, SAP, SAS, Birst, Yellowfin and ElegantJ BI are all examples of this category.

Category No. 4: Integrated as part of a data science/machine-learning tool. These integrated data preparation vendor offerings focus on data preparation capabilities as part of an end-to-end data science and machine-learning process and offering, with broader advanced analytics capabilities. Many data science and machine-learning platforms have integrated data preparation and data pipelining features (see "Magic Quadrant for Data Science Platforms"). Alteryx, IBM, Lavastorm, RapidMiner, Rapid Insight and SAS are examples.

Note: While we highlight the sample embedded vendors (in Categories No. 2, 3 and 4) in Table 1, the write-ups/analysis focus only on stand-alone data preparation vendor offerings (Category No. 1).

Data Preparation Vendors by Category

	Stand-Alone Data Preparation Tool	Integrated as Part of a Data Integration Tool	Integrated as Part of a Modern Analytics and BI Platform	Integrated as Part of a Data Science/Machine-Learning Platform
Alteryx	✓		✓	✓
Cambridge Semantics	✓		✓	
ClearStory Data	✓		✓	
Datameer	✓			
Datawatch	✓		✓	
IBM	✓		✓	✓

Data Preparation Vendors by Category

	Stand-Alone Data Preparation Tool	**Integrated as Part of a Data Integration Tool**	**Integrated as Part of a Modern Analytics and BI Platform**	**Integrated as Part of a Data Science/Machine-Learning Platform**
Lavastorm	✓			✓
Lore IO	✓	✓		
Oracle	✓		✓	
Paxata	✓			
Rapid Insight	✓			✓
SAP	✓		✓	
SAS	✓	✓	✓	✓
Talend	✓	✓		
Tamr	✓			
TMMData	✓	✓	✓	
Trifacta	✓			
Unifi	✓			
ElegantJ BI			✓	
Hitachi Vantara (formerly Pentaho)		✓	✓	
Informatica		✓		
Microsoft			✓	
MicroStrategy			✓	
Podium Data		✓		
Qlik			✓	
Tableau			✓	
Yellowfin			✓	
Zaloni		✓		

Data Preparation Vendors by Category

	Stand-Alone Data Preparation Tool	Integrated as Part of a Data Integration Tool	Integrated as Part of a Modern Analytics and BI Platform	Integrated as Part of a Data Science/Machine-Learning Platform

✓= Vendor belongs to the particular segment/category of data preparation