Market Guide for Data Preparation Tools

Published 9 July 2020 - ID G00719343 - 47 min read

By Analysts Ehtisham Zaidi, Sharat Menon

Data preparation is a must-have technology for enabling business teams to find, prepare and share heterogeneous data for their integration, analytics and data science use cases. Data and analytics leaders must understand the dynamics of this evolving market to identify suitable vendor offerings.

Overview

Key Findings

- Data preparation tools have evolved from being able to support only self-service use cases, and now enable data and analytics teams to build integrated datasets at an enterprise scale in an agile manner for a range of distributed content authors.
- The market for data preparation tools remains crowded and complex. Choices range from stand-alone specialists to vendors that embed data preparation as a key capability into their broader analytics/BI, data science/ML, cloud ecosystems or data integration tools.
- While most data preparation tool capabilities have been maturing at a steady state, organizations continue to cite "operationalization" — the ability to promote self-service models to production environments — and "augmented data preparation" — i.e., the utilization of embedded ML/AI algorithms to simplify (and in some cases automate) data preparation — as the two biggest differentiators that enable enterprisewide adoption.

Recommendations

Data and analytics leaders focused on data management solutions should:

- Deploy data preparation tools strategically with a focus on enhancing user understanding of data and reducing data preparation efforts for increased productivity. Prioritize adoption of those data preparation tools that automate the process of moving self-service workloads into production environments through metadata and integration support.
- Evaluate stand-alone data preparation tools when your use case is a general-purpose one needing preparation of data for analysis across multiple analytics/BI, data science/ML, integration and CSP platforms. Conversely, evaluate the embedded data preparation capability of incumbent tools if you need data preparation only in the context of those tools, platforms or ecosystems.
- Evaluate data preparation tools for their ability to scale from self-service models to enterprise-level projects. Give preference to tools that can coexist with other data management tools (such as data catalogs, data quality and information stewardship) and have the ability to capture, analyze and share metadata (and lineage) with them, to ensure security, governance and compliance practices.

Strategic Planning Assumption(s)

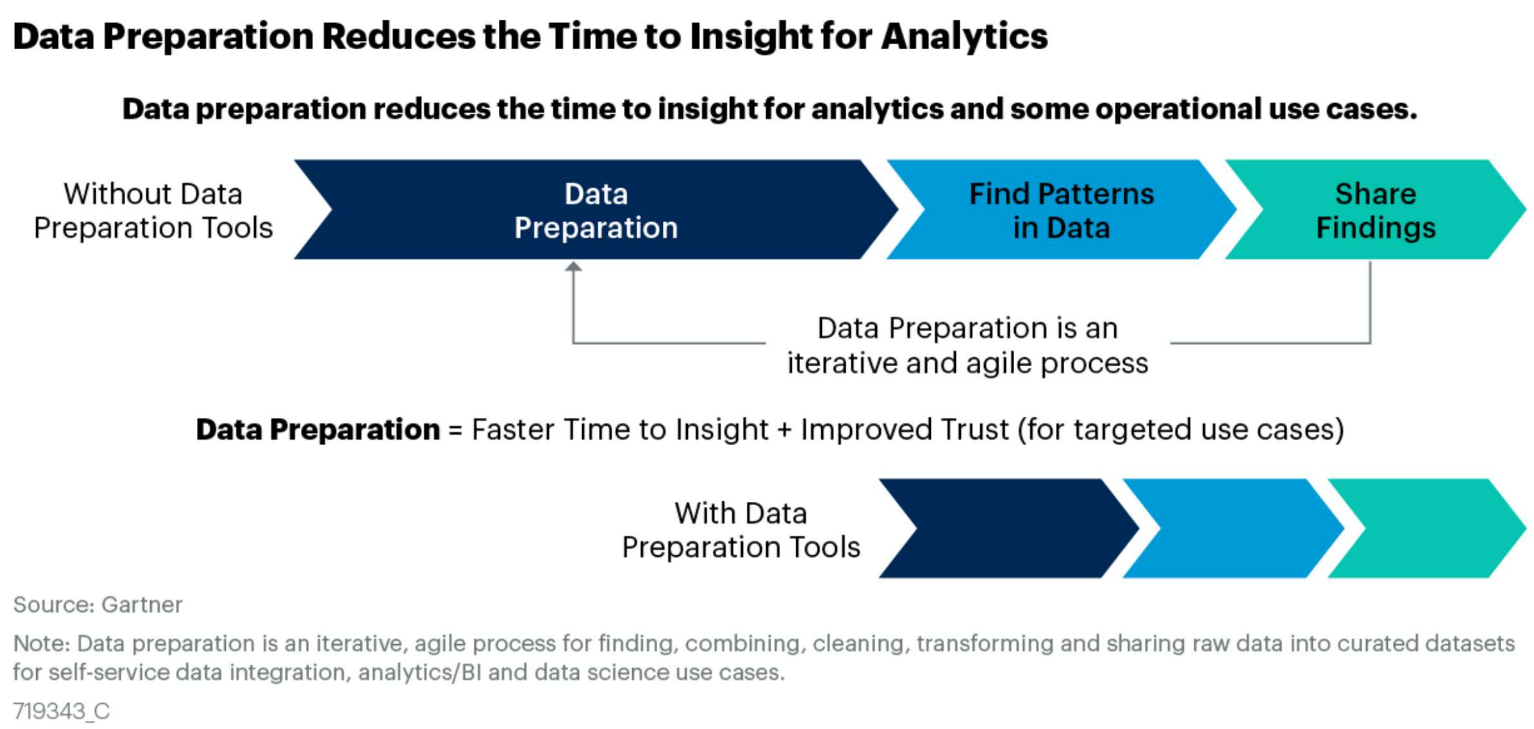
- By 2022, data preparation will become a critical capability in more than 80% of data integration, analytics/BI, data science, data engineering and data lake enablement platforms.
- By 2021, organizations that offer users access to a curated catalog of internal and external prepared data will realize twice the business value from analytics investments than those that do not.
- By 2024, augmented data preparation, data catalogs, data unification, data virtualization and data quality tools will converge into a consolidated data fabric used for the majority of new analytics/data science projects.

Market Definition

Data preparation is an iterative and agile process for finding, combining, cleaning, transforming and sharing curated datasets for various data and analytics use cases including analytics/business intelligence (BI), data science/machine learning (ML) and self-service data integration.

Data preparation tools promise faster time to delivery of integrated and curated data by allowing business users including analysts, citizen integrators, data engineers and citizen data scientists to integrate internal and external datasets for their use cases. Furthermore, they allow users to identify anomalies and patterns and improve and review the data quality of their findings in a repeatable fashion. Some tools embed ML algorithms that augment and, in some cases, completely automate certain repeatable and mundane data preparation tasks. Reduced time to delivery of data and insight is at the heart of this market (see Figure 1).

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



Market Description

The data preparation tool market consists of stand-alone data preparation tools as well as tools from other data and analytics markets that provide data preparation as an embedded critical capability.

This market caters primarily to those users who were previously left frustrated by the slow turnaround time for IT developers or data engineers to provide integrated and curated data ready for analytics, data science and other operational use cases. The promise of data preparation is to increase the productivity of data users by reducing the time that they spend on data preparation tasks and giving them that time back for performing more productive tasks, such as business analytics and data-science-related activities. In order to achieve the twin goals of faster time to insight and improved trust, data preparation tools should have certain key functionalities or capabilities (see Figure 2).

Figure 2: Key Components of a Modern Data Preparation Tool

Key Components of a Modern Data Preparation Tool

Source: Gartner
719343_C

For a brief description of these capabilities and a summary of how data preparation tools stack up across them, see "Tool: Evaluate Data Preparation Tools Across Key Capabilities" and the descriptions below:

- **Data source access/connectivity —** Breadth of connectivity options like APIs and native access to cloud data sources (such as popular database platform as a service [dbPaaS] and cloud data warehouses), on-premises databases, SaaS applications, nonrelational databases (including Hadoop and cloud object stores), traditional and modern file formats, connectivity to mobile platforms, event/streaming data, as well as native access to open, premium or curated data.
- **Data profiling and data quality support —** Users increasingly demand a visual environment that enables users to interactively profile, search, sample and prepare data assets, and to tag, link, match, merge, deduplicate and annotate data for future exploration.
- **Data cataloging —** Users of data preparation tools demand embedded data cataloging capabilities that allow them to create and search for metadata, publish curated content, catalog key data sources, catalog repeatable transformations and user activity against the data source, data source attributes, data lineage and relationships, and APIs.
- **Data modeling and transformation —** Support for agile data modeling/structuring that allows users to specify data types and relationships. Advanced capabilities include the ability to automatically deduce or infer the structure from the data source and generate semantic models and ontologies — such as logical data models and schemas.

- **Data security —** Inclusion of security features such as data masking, platform authentication and security filtering at the user/group/role level through integration with corporate Lightweight Directory Access Protocol (LDAP) and/or Active Directory (AD) systems, single sign-on (SSO), source system security inheritance, and row- and column-level security.

- **Basic data governance/information stewardship support —** Support for a bidirectional metadata exchange and integration with broader data governance/information stewardship tools. This would provide capabilities for tracking data lineage, providing impact analysis reports, enforcing policies and adhering to user permissions.

- **Data enrichment —** Support for data enrichment capabilities including entity extraction, capture of attributes from the integrated data using the data preparation tool. In addition, attribute development, which enables process experts to develop the attribute set for integrated data based on the requirements of their industry or domain.

- **User collaboration and operationalization —** The ability to share queries and datasets, including publishing, sharing and promoting developed models (from sandbox to production environments) with governance features such as dataset user ratings or official watermarking.

- **API-based access to downstream data consumption applications —** The ability to share the developed models to downstream tools for analysis, sharing and consumption. Data preparation tools must provide API-based access to popular analytics/BI, data science/ML and other message queues for data sharing and analysis.

- **Augmented data preparation (for insights and automation) —** This differentiating capability augments data preparation tools through embedded ML/AI algorithms to improve and, in some cases, even automate the data preparation process. This includes algorithms that enable users to identify data structures, schemas and relationships, and the ability to structure the datasets upon initial data ingestion. These algorithms over time recommend, with more precision, the most accurate joins, transformations and enrichments. Some tools even automate data tagging, profiling, anomaly detection (and reporting), and detection of sensitive data and risky attributes (such as PII data).

- **Hybrid and multicloud deployment options —** Organizations need data preparation tools that can be deployed either in the cloud (through true platform as a service [PaaS] deployments), on-premises, or in a hybrid integration platform setting. The hybrid approach allows users to leave data in place (for processing), rather than having to replicate it in the data preparation tools engine — reducing data silos.

- **Domain- or vertical-specific offerings or templates —** Packaged templates for domain- or vertical-specific data and models that can accelerate the time to data preparation. This is helpful for several syndicated datasets that are otherwise difficult to prepare and model. Such packaged offerings could include domain accelerators, industry starter templates, best practices, KPIs and vertical/domain-specific IP that support industry specific models such as HL7, SWIFT and HIPAA, among several others.

Market Direction

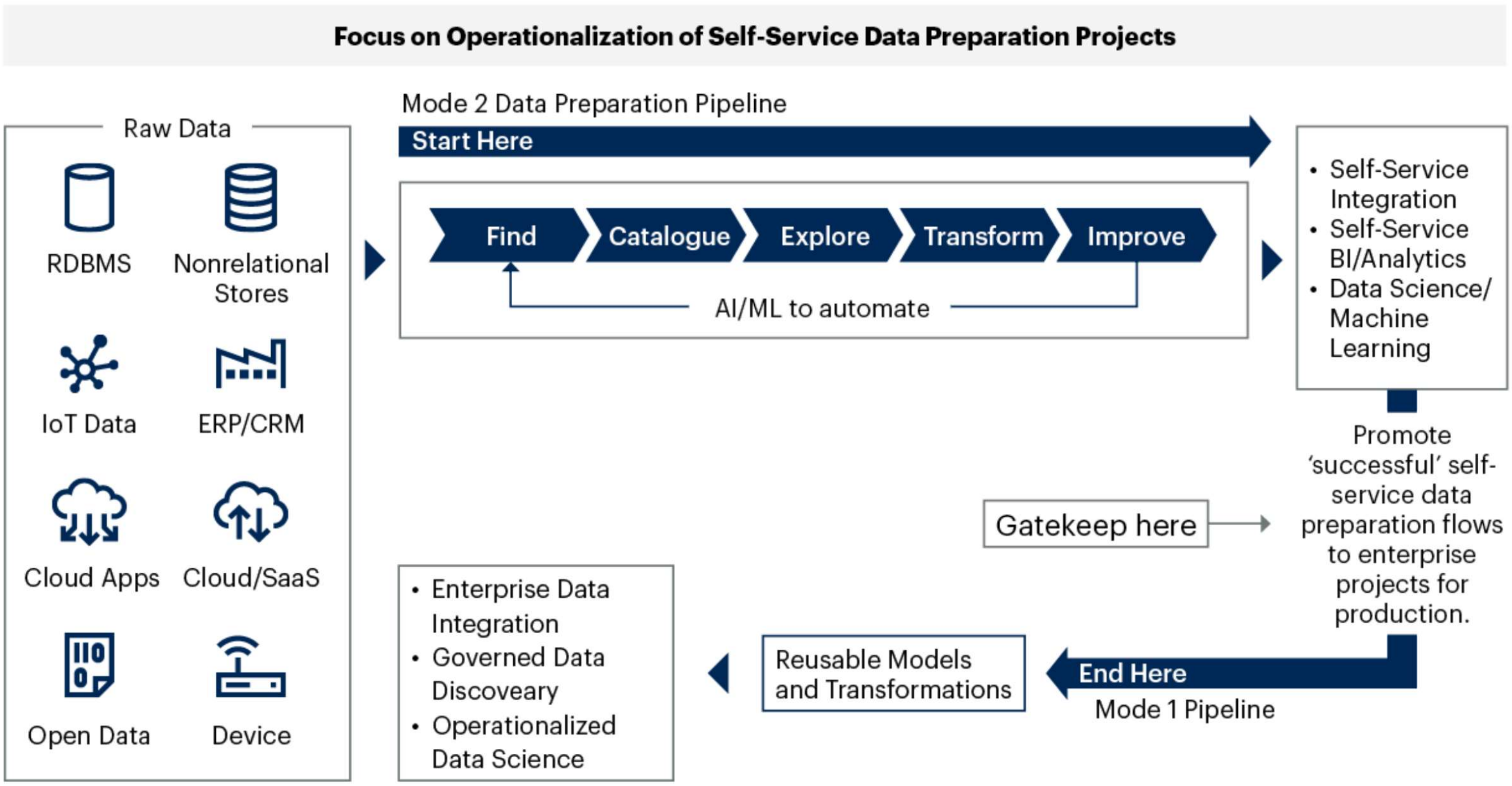
There are a number of trends that will continue to drive adoption of data preparation tools in data and analytics markets, cutting across the behavior of both buyers and vendors.

Trends in Buyer Behavior

- *Operationalization of data preparation flows continues to be the most important differentiator.* Most data preparation tasks start as self-service (Mode 2) activities in the data preparation pipeline (see Figure 3). This is where business analysts, citizen data scientists, citizen integrators and other users of data preparation tools connect to multiple, frequently changing data sources to prepare the data for experimentation. Once the Mode 2 experiment is successful and the business wants to operationalize its findings (promote to a system of record), it must ensure that IT/data engineering teams provide the required promotion processes based on data quality and governance strategies. Once the quality and governance of the prepared datasets are certified, data preparation tools must be able to facilitate or even automate the promotion of these self-service data preparation flows to broader and reusable Mode 1 transformations.

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

Operationalization of Data Preparation — The Modern Data Preparation Pipeline



Source: Gartner
719343_C

- *Increased expectation (and use) of machine learning to support augmented data preparation.* Embedded ML capabilities are now a "must-have" in data preparation tools. Buyers seek tools that can provide out-of-the-box augmented capabilities for automated matching, joining, profiling, tagging and annotating connected data prior to data preparation. Advanced capabilities include highlighting sensitive attributes, anomalies and outliers, and collaborating with metadata management and governance tools to prevent sensitive data from being exposed. Buyers also look for ML capabilities to inform and even automate repetitive transformations and integrations, to assist less-skilled users to perform integration tasks. For further details, see "Augmented Analytics Is the Future of Analytics" and "Rebalance Your Integration Effort With a Mix of Human and Artificial Intelligence."

- *Increased requirement for "cloud/as a service, hybrid and multicloud" options.* Buyers are looking for true PaaS options for their data preparation tools, which provide the flexibility and scalability of cloud data integration. They are looking for tooling that can provide flexibility in scaling storage and compute resources, when needed, to toggle between self-service and complex/enterprise-level data workloads. Few vendors today offer the ability to take the processing to the data, and vice versa, across multicloud and hybrid cloud options. Buyers must look at their prospective vendors' roadmaps to ensure that their tools can prepare data across a hybrid cloud (on-premises and cloud) and multicloud (across different CSPs), to prevent cloud lock-in.

- *Data lakes need data preparation capabilities.* Popular data lake enablement platforms, used in the capacity of a data science lab, can benefit from data preparation tools to empower non-IT users to prepare data for analytics or data science needs (see "Use Design Patterns to Increase the Value of Your Data Lake"). Gartner is seeing most data lake enablement offerings providing data preparation as a critical capability embedded within their platform, either through their own data preparation tool or through partnerships with stand-alone data preparation tool vendors.

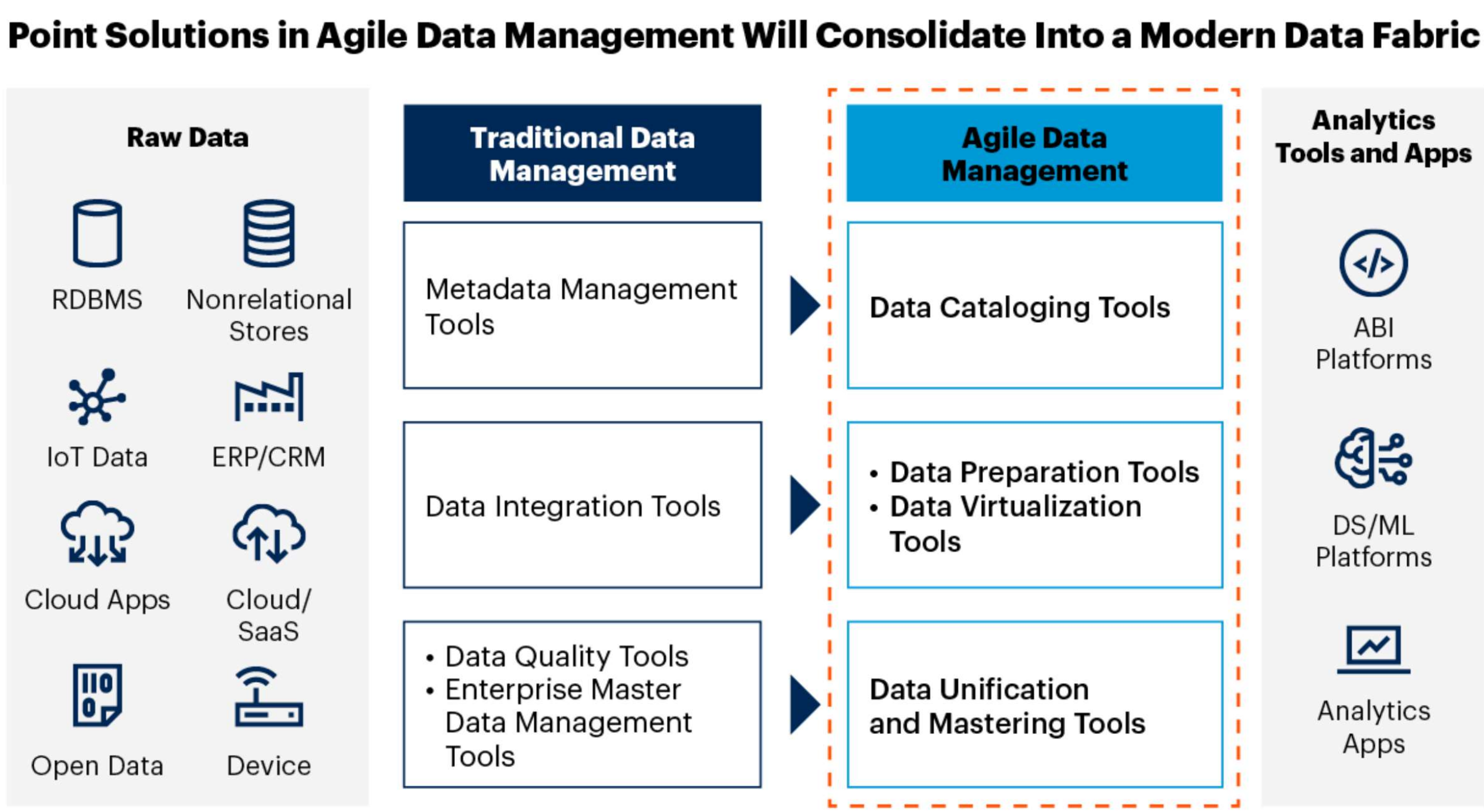
- *Support for data engineers.* Data engineers are in short supply and are tasked with the cumbersome role of building and managing data pipelines and promoting these data pipelines to production (in line with business processes). They must therefore be provided with data preparation tools to assist them in building these data pipelines faster, with minimal coding, thereby making them more productive (see "Toolkit: Job Description for the Role of a Data Engineer").

Market Analysis

Trends in Vendor Behavior and Overall Market Outlook

- **Data preparation tools along with multiple other point solutions will rapidly converge into a consolidated data fabric.** A data fabric design requires capabilities for business-led data modeling and schema and semantics assignment. This is why data preparation capabilities are becoming an important part of the overall data fabric architecture. Going forward, Gartner expects various business-oriented data management tools to converge into a data fabric supporting broader data management use cases. The key capabilities that we see converging include data catalogs, data preparation, data unification (for agile mastering of data) and data virtualization (for agile integrated data access and delivery). These converging data management capabilities are targeted toward business users (see dark blue boxes in Figure 4 below) as a direct response to similar data management tools that were mostly targeted toward skilled IT users (see light blue boxes in Figure 4 below). Gartner expects these two converged data management platforms (one focused on IT users and one focused on business users) to coexist, and the data fabric will be the architecture that enables and supports this coexistence (see "Data Fabrics Add Augmented Intelligence to Modernize Your Data Integration").

Figure 4: Point Solutions Will Consolidate Into a Modern Data Fabric



Source: Gartner
719343_C

- **Inorganic growth will continue for the next two to three years, leading to eventual consolidation; however, there is still room for a stand-alone data preparation tool for specific use cases.** Vendors from adjacent markets — such as data integration, analytics/BI, data science/ML and others — are acquiring stand-alone data preparation vendors to quickly ramp up their data preparation capabilities. For example, Unifi Software, a provider of data preparation and data catalog solutions, was acquired by Boomi, an enterprise integration platform as a service (iPaaS) provider, to provision data preparation and catalog capabilities as a cloud delivery model. Paxata, one of the leading data preparation vendors in the market, was acquired by DataRobot, a data science vendor, to embed data preparation as a capability for data engineers and data scientists working on AI/ML use cases. Lavastorm was acquired by Infogix, and Podium Data was acquired by Qlik.

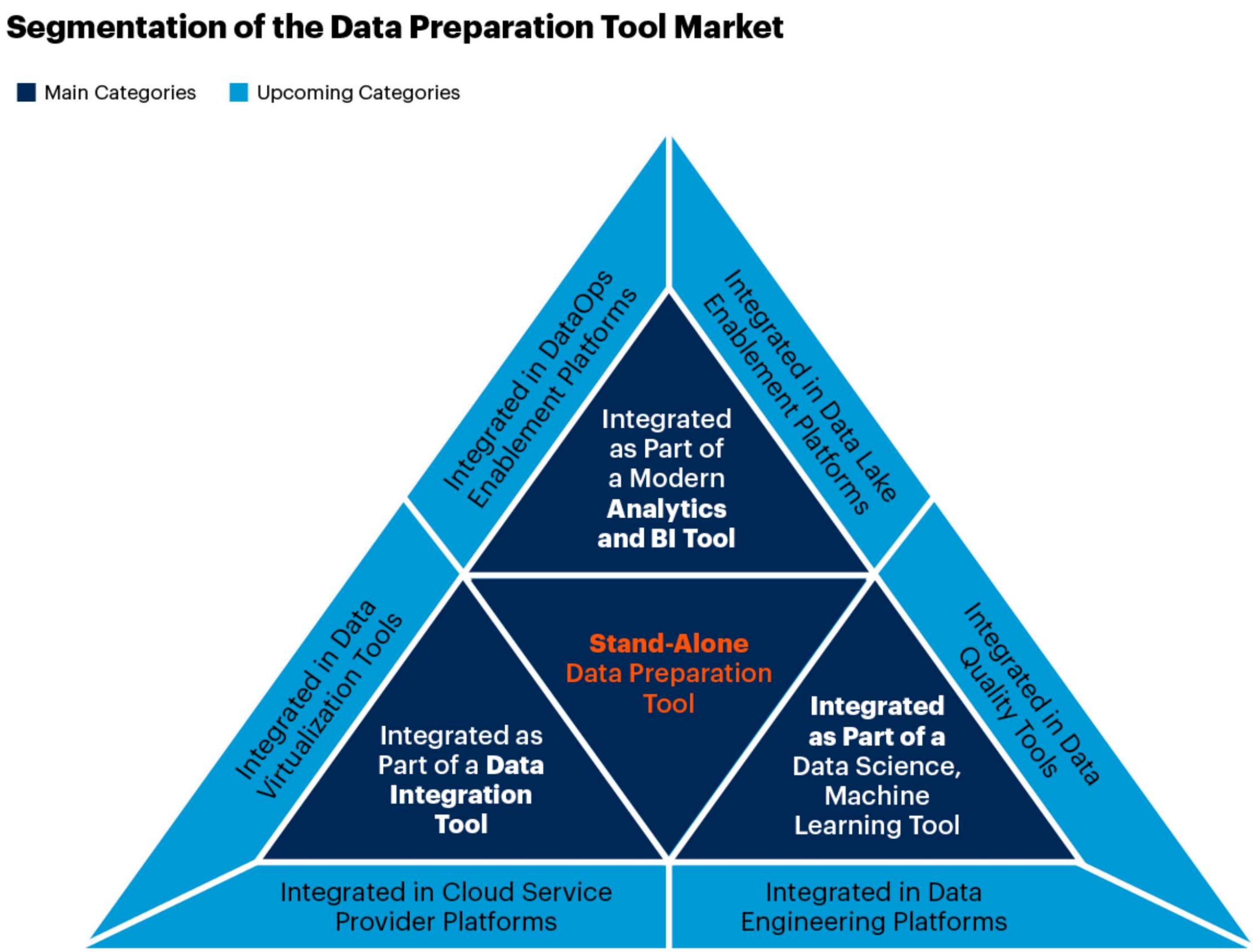
On the other hand, vendors that offer stand-alone data preparation tools are also making key acquisitions to complement and extend their data preparation tools to include end-to-end analytics and data management capabilities. For example, Alteryx recently acquired ClearStory Data (a continuous intelligence analytics solution focused on intelligent automation on large-scale data processing platforms). Datawatch (now Altair) acquired Angoss, to complement its data preparation capabilities with predictive analytics. Acquisitions are expected to continue in the next two to three years, for vendors to augment and complement their capabilities.

However, there will still be room for stand-alone data preparation tools that offer highly specialized and use case/domain/vertical industry-focused data preparation capabilities — such as data preparation for healthcare industry data, or data preparation for insurance domain with quick access to external open datasets. There is also a current gap in the market for data preparation tools with the ability to handle highly specialized or complex data preparation use cases very well. These use cases include Internet of Things (IoT) data preparation, stream data preparation and location or geospatial data preparation (see "Adopt Stream Data Integration to Meet Your Real-Time Data Integration and Analytics Requirements").

Main Segments of the Data Preparation Tool Market

This Market Guide highlights data preparation vendors that are segmented into four broad categories (dark blue) and six upcoming categories (light blue) (see Figure 5 below).

Figure 5: Segmentation of the Data Preparation Tool Market



Source: Gartner
719343_C

Category 1. Stand-Alone Data Preparation Tool

Vendors in this category sell data preparation as a stand-alone, independent or pure-play offering. Stand-alone vendor offerings focus on enabling tighter integration with downstream processes, such as support for multiple analytics/BI, data science and data integration tools. They are independent, so that buyers can buy them specifically for their data preparation requirements, and they support data preparation for general-purpose use-case requirements — such as supporting data preparation for downstream analysis in multiple analytics/BI or data science platforms/tools.

These stand-alone data preparation vendors are the focus of this Market Guide and have been listed first in Table 1 (above the black line separator) and analyzed in detail in the Representative Vendors section.

Category 2. Integrated as Part of a Data Integration Tool

Vendors here are focused on data integration and management, and have added data preparation to their product portfolios. This is done either by embedding data preparation capabilities into their existing data integration tool portfolio, or as a separate data preparation tool module that can be purchased to support their data integration/data management 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 Adeptia, Ataccama, Denodo, Infogix, Informatica, Microsoft, Oracle, Qlik (through its acquisition of Podium Data — provides Qlik Data Catalyst), SAP, SAS, Talend and Boomi (through their acquisition of Unifi Software) already include data preparation in their enterprise data integration and data management portfolios.

Category 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 workflow, with broader analytics and BI and content creation capabilities. All modern analytics and BI vendors have some embedded data preparation capability as this is a critical capability (see "Magic Quadrant for Analytics and Business Intelligence Platforms" and "Other Vendors to Consider for Modern Analytics and BI" for all vendors with this capability). Alteryx, Tableau, Elegant MicroWeb, Infogix, Microsoft, MicroStrategy, Oracle, Qlik, SAP, SAS and TIBCO Software are all examples of this category.

Category 4. Integrated as Part of a Data Science and Machine Learning Platform

These integrated data preparation vendor solutions focus on data preparation capabilities as part of an end-to-end data science and ML process and offering, with broader advanced analytics capabilities. Many data science and ML platforms have integrated data preparation and data pipelining features (see "Magic Quadrant for Data Science and Machine Learning Platforms"). Alteryx, Dataiku, DataRobot (through its acquisition of Paxata), Explorium, IBM, Infogix, Rapid Insight, SAP and SAS are all examples.

Upcoming Categories

With data preparation becoming a key capability across all data management and analytics tools, Gartner is now seeing data preparation capabilities being embedded across several new categories:

- **Data preparation as an embedded capability in cloud service providers (CSPs)** — all popular CSPs now include data preparation capabilities as a part of their broader data and analytics platform portfolio. Examples include IBM Cloud (IBM InfoSphere Advanced Data Preparation, through a partnership with Trifacta), Microsoft Azure (with data preparation capabilities delivered as part of its Power BI tooling) and Google Cloud Platform [GCP] (through its partnership with Trifacta and its embedded data preparation capabilities in Google Cloud Data Fusion).

- **Data preparation embedded as a capability in broader data lake enablement, DataOps and data engineering platforms** — Data preparation is now being included as a capability in upcoming platforms for data engineering/stream data integration, DataOps and data lake enablement. Common examples of this include Infoworks (DataFoundry), Nexla (Data Operations), StreamSets (StreamSets DataOps Platform), Striim (The Striim Platform) and Zaloni (Zaloni Arena Platform) (see "Data Engineering Is Critical to Driving Data and Analytics Success," "Innovation Insight for DataOps," and "Adopt Stream Data Integration to Meet Your Real-Time Data Integration and Analytics Requirements").

- **Data preparation embedded as a capability in data quality tools —** Data preparation is included as a key capability in data quality tools such as Ataccama (Ataccama ONE), Experian (Aperture Data Studio) and others (see "Magic Quadrant for Data Quality Tools").

- **Data preparation as an embedded capability in data virtualization tools —** Common examples of data virtualization tools that embed some data preparation capabilities include Denodo (Denodo platform), TIBCO (TIBCO Data Virtualization), Dremio (Dremio Enterprise Edition) and others (see "Market Guide for Data Virtualization").

Note that, while we highlight the sample embedded vendors (listed in Categories 2, 3 and 4 above) in Table 1, however, the write-ups/analyses in the Representative Vendors section focus only on the stand-alone data preparation vendor offerings (Category 1), which are listed first (above the black line separator) in Table 1.

Representative Vendors

The vendors listed in this Market Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

Market Introduction

The vendors and products listed and analyzed in this section are representative because they have achieved some level of visibility and traction in this market (see Note 1). Vendors are widely diverse in their capabilities, although all support the general template described earlier in the Market Definition and Market Description sections.

Table 1 below lists the profiled vendor tools categorized according to the data preparation segments described in the Market Analysis section. The lists are not intended to be comprehensive, but rather representative of the market. Please note that vendors that offer at least one stand-alone data preparation tool are listed first in Table 1 (in alphabetical order above the table break); those that provide data preparation embedded only as part of the other categories (rather than stand-alone) are listed next (after the table break shown by a black separator in Table 1).

While many vendors could be offering data preparation capabilities embedded within their broader tools/platforms, we list only those about which Gartner has received the most client interest (according to searches on gartner.com and our internal client inquiry service).

For more analysis on data preparation tools and their capabilities, typically for tool evaluation and selection, see "Tool: Evaluate Data Preparation Tools Across Key Capabilities."

Table 1: Details of Vendors (and Tool Names) Providing Data Preparation

Vendor Name (top to bottom) ↓	*Tool Names Across Different Categories* ↓			
	Stand-Alone Data Preparation Tool	**Data Preparation Capability Embedded Within Data Integration Tool Or Data Management Platform**	**Data Preparation Capability Embedded Within Modern Analytics/BI Platform**	**Data Preparation Capability Embedded Within Data Science/ML Platform**
Altair (formerly Datawatch)	Altair Monarch, Altair Knowledge Hub, Altair Knowledge Works	-	-	Altair Knowledge Studio
Alteryx	Alteryx Designer	-	Alteryx Analytic Process Automation (APA) platform	Alteryx Analytic Process Automation (APA) platform
Boomi (previously Unifi Software)	Boomi Data Catalog and Preparation (DCP)	Boomi Atomsphere platform	-	-
Datameer	Datameer X, Neebo	-	-	-
DataRobot-Paxata	Paxata Adaptive Integration Platform	-	-	-
Elegant MicroWeb	Smarten Self-Serve Data Preparation	-	Smarten Augmented Analytics	Smarten Insight

Vendor Name (top to bottom) ↓	*Tool Names Across Different Categories* ↓			
Explorium	Explorium Data Science Platform			Explorium Data Science Platform
Google Cloud Platform (GCP)	Cloud Dataprep by Trifacta (a cloud service to support data preparation across data management,data science and analytics use cases on GCP)			
IBM	InfoSphere Advanced Data Preparation (through partnership with Trifacta)	Watson Knowledge Catalog service in IBM Cloud Pak for Data	Watson Knowledge Catalog service in IBM Cloud Pak for Data	Watson Knowledge Catalog service in IBM Cloud Pak for Data
Infogix	Data360 Analyze	Data360 Govern, Data360 DQ+	Data360 Analyze	-
Informatica	Enterprise Data Preparation	Enterprise Data Preparation (for broader data integration and data management)	-	-
Lore IO	Lore IO	-	-	-
Modak	Modak Nabu	Modak Nabu	-	-
Quest Software	Toad Data Point	-	-	-
Rapid Insight	Rapid Insight Construct	-	Rapid Insight Bridge	Rapid Insight Predict
SAP	SAP Agile Data Preparation	SAP Data Intelligence	SAP Analytics Cloud	SAP Data Intelligence
SAS	SAS Data Preparation (on SAS Viya), SAS Data Loader for Hadoop	SAS Data Loader for Hadoop (for broader data integration and data management)	SAS Data Preparation (on SAS Viya)	SAS Data Preparation (on SAS Viya)
Talend	Talend Data Preparation	Talend Data Fabric (for broader data integration and data management)	-	-
Tamr	Tamr Enterprise Data Mastering (for data mastering and data unification)	-	-	-

Vendor Name (top to bottom)		*Tool Names Across Different Categories*		
Trifacta	Trifacta Editions include Free, Standard, Premium and Self-Managed.	-	-	-
Adeptia	-	Adeptia Connect (for data integration)	-	-
Ataccama	-	Ataccama ONE (for data integration, data quality and broader data management)	-	-
Dataiku	-	-	-	Data Science Studio (DSS)
Denodo	-	Denodo Platform (for data integration and data virtualization)	-	-
Dremio	-	Dremio Enterprise Edition (for data integration and virtualization)	-	-
Experian	-	Aperture Data Studio (for data quality)	-	-
Infoworks	-	Infoworks DataFoundry (for DataOps)	-	-
KNIME	-	-	-	KNIME Analytics Platform
Microsoft	-	SQL Server Integration Services (SSIS) and Azure Data Factory (by utilizing Power Query's data preparation capability) (for data integration and broader data management)	Power BI (by embedding and utilizing the data preparation capabilities of Power Query, Power Pivot and Power View)	-
MicroStrategy	-	-	MicroStrategy	-
Nexla	-	Nexla Data Operations (for DataOps)	-	-
Oracle	-	Oracle Cloud Infrastructure Data Integration (for data integration and broader data management)	Oracle Analytics Cloud	-
Qlik	-	Qlik Catalog (for data integration and broader data management)	Qlik Sense	-
Radicalbit	-	RNA (for stream data integration and DataOps)	-	-

Vendor Name (top to bottom)	Tool Names Across Different Categories			
StreamSets	-	StreamSets DataOps Platform (for stream data integration and DataOps)	-	-
Striim	-	The Striim platform (for stream data integration)	-	-
Tableau	-	-	Tableau Prep — composed of Tableau Prep Builder (available stand-alone) and Tableau Prep Conductor (available as part of the Tableau Data Management Add-on)	-
TIBCO Software	-	TIBCO Data Virtualization (for data integration and virtualization)	TIBCO Spotfire	-
Zaloni	-	Zaloni Arena (for DataOps)	-	-
(-) sign = not applicable; vendor does not belong to the segment/category				

Source: Gartner (July 2020)

The following write-ups profile and analyze only the vendors listed first in Table 1 (above the black line separator) — those that offer at least one stand-alone data preparation tool.

The vendors listed in this Market Guide do not imply an exhaustive list. This section is intended to provide more understanding of the market and its offerings.

Vendor Profiles

Altair (formerly Datawatch)

www.altair.com

Product name: Altair Monarch, Altair Knowledge Hub, Altair Knowledge Works

Headquarters: Troy, Michigan, U.S.

Geographic presence: Asia/Pacific, EMEA, North America, Latin America

Licensing model: Altair Monarch runs on desktops and is licensed by named users as a subscription or as a perpetual license. The Monarch Automation Server is priced per process, starting with 25 processes. Altair Knowledge Hub is sold on a subscription-only, server- and user-based pricing model. It offers a creator license and a consumer license.

Number of deployments: Not provided by the vendor

Altair Knowledge Hub is a browser-based data preparation solution, while Altair Monarch is a desktop-based data preparation tool. The key differentiator is the ability to extract data via web scraping and from semistructured data sources such as Excel, PDF files and text files. Altair provides interoperability between these two solutions. For instance, a business user can export data from Monarch into Knowledge Hub for operationalization. Critical capabilities, such as a data catalog including trustworthiness ranking of cataloged assets, data lineage, security (AD, LDAP integration, role-based access control, etc.) and deployment on public cloud and via Kubernetes, make this a well-rounded set of offerings.

Alteryx

www.alteryx.com

Product name: Alteryx Designer, Alteryx Analytic Process Automation platform

Headquarters: Irvine, California, U.S.

Geographic presence: North America, Latin America, EMEA and Asia/Pacific

Licensing model: Licensed use on an annual subscription basis, and priced per named user

Number of deployments: 6,400 customers

Alteryx Designer is a stand-alone product that provides a low code interface for exploring and preparing data for data science use cases, such as predictive, prescriptive and geospatial analytics, for citizen data scientists. The Alteryx Intelligence Suite is a Python-based add-on to Designer that provides assisted modeling via augmented machine learning, as well as the ability to analyze semistructured and unstructured data. This is typically used by data scientists. Clients therefore view Alteryx as a viable solution when they do not want to invest in multiple solutions for code-intensive data science (R or Python code can be integrated into an Alteryx workflow) and code-free data preparation. The Alteryx Analytics Hub is used to automate and operationalize Designer workflows in a central, secure, governed analytics environment. All these components come together in a unified platform called the Alteryx Analytic Process Automation platform. Data can be processed in place, and in massively parallel processing (MPP) engines such as Apache Avro, Apache Spark and Databricks. The server-based components of the platform can be deployed on Microsoft Azure, Amazon Web Services (AWS) and Google Cloud Platform.

Boomi (formerly Unifi Software)

boomi.com

Product name: Boomi Data Catalog and Preparation (DCP)

Headquarters: Chesterbrook, Pennsylvania, U.S.

Geographic presence: North America, EMEA, Asia/Pacific

Licensing model:

- Data Catalog only: Base fee + five Catalog only users. Additional tiers of Data Catalog user licenses can be purchased at an additional "per user" cost/tier.
- Data Catalog and Preparation: Three-tier base fee includes (x number of processing nodes/tier, y number of catalog users/tier, z number of prep users/tier).
- Additional Data Prep and Catalog users/tier can be added for additional cost.

Number of deployments: Not provided by the vendor

Boomi, a Dell Technologies business, announced its intent to acquire Unifi Software on 17 December 2019. Boomi broadly markets Unifi Data Platform (UDP), now rebranded as Boomi Data Catalog and Preparation (DCP). Customers of Boomi can utilize DCP to catalog and prepare their data before utilizing Boomi Atomsphere Platform for their broader integration use cases such as iPaaS, MDM and B2B data sharing.

DCP provides augmented data preparation and cataloging in a single platform. DCP's data cataloging capability enables users to find data using natural language queries. DCP's data exploration capabilities include automated data parsing and metadata search, automated profiling, cleansing and object indexing (using its one-click functions), and augmented recommendations on datasets, join types, filtering and formatting (using its OneMind feature). Support for data governance is provided through role-based access, and security is implemented through AD, Kerberos, SAML, row- and column-level security, and so on. MPP support comes from Spark and Impala, and a cost-based optimizer is leveraged to select the most appropriate environment for processing the data transformations. For cloud deployment, Boomi's DCP is available on AWS, Microsoft Azure, GCP and Adobe Experience Cloud.

Datameer

www.datameer.com

Product name: Datameer X, Neebo

Headquarters: San Francisco, California, U.S.

Geographic presence: Asia/Pacific, EMEA, North America, South America

Licensing model: Customers are charged based on compute power (V-Cores) or data volume. Cloud customers are charged hourly or via an annual license.

Number of deployments: 120

Datameer X is a data platform that supports agile, scalable DataOps, and provides tools for data preparation, data exploration and operationalization of data pipelines within the platform. Datameer X has a Kubernetes-based container architecture powered by Spark that is elastic (autoscaling) and supports cloud, hybrid and on-premises deployments. It also supports multiple scale-out and performance optimization techniques and push-down into MPP and cloud data warehouses. The addition of Neebo to the product portfolio enables Datameer to prepare data in place, without the need for data movement and replication processes.

DataRobot-Paxata

www.paxata.com

Product name: Paxata Self-Service Data Preparation

Headquarters: Redwood City, California, U.S.

Geographic presence: Asia/Pacific, EMEA, Latin America, North America

Licensing model: Annual, subscription-based license model

Number of deployments: Not provided by the vendor

Paxata Self-Service Data Preparation (SSDP), which is both a stand-alone solution for data preparation and part of the Paxata Adaptive Information Platform, provides excellent ML-embedded automation and NLP support as its core differentiators. Customers can easily switch between batch and interactive workloads using Paxata Adaptive Workload Management. Data catalog support is provided through partnerships with Alation and Waterline Data. Its acquisition by DataRobot will expand its customer pool for the data science use case.

Elegant MicroWeb

www.smarten.com

Product name: Smarten

Headquarters: Ahmedabad, India

Geographic presence: Asia/Pacific

Licensing model: Two licensing models are available — a user-based license, and an Enterprise license with core-based pricing.

Number of deployments: More than 80

Smarten is a browser-based augmented analytics solution with an embedded data preparation module — Self Serve Data Preparation (SSDP) — that can also be purchased separately. Support for data governance is provided through access rights management (including column-level access control), data lineage with action rollback capabilities and IT certification of data assets. Other adjacent capabilities provided are clickless search analytics using NLP, support for out-of-the-box predictive models and visual data discovery (powered by another module called Smarten View). For public cloud options, Smarten can be deployed on AWS and is available on AWS Marketplace.

Explorium

www.explorium.ai

Product name: Explorium Data Science Platform

Headquarters: San Mateo, California, U.S.

Geographic presence: North America, EMEA, Asia/Pacific

License model: Annual subscription SaaS model

Number of deployments: Not provided by vendor

Explorium is a data discovery and feature generation platform that automates the end-to-end process of data acquisition, preparation, and transformation to drive business insights. The platform can be used as a stand-alone data preparation solution and uses advanced algorithms to automatically connect the user to thousands of external data sources (premium, partner, and open), merge and match those external sources with their internal data, and automatically distill and rank top features. This allows data teams to improve the accuracy of their models with data enrichment, and prepare, assemble and deploy models for superior business outcomes.

Google Cloud Platform (GCP)

cloud.google.com

Product name: Cloud Dataprep by Trifacta

Headquarters: Mountain View, California, U.S.

Geographic presence: North America, EMEA, Asia/Pacific, Latin America

Licensing model: Priced by number of virtual CPUs

Number of deployments: Not provided by the vendor

Cloud Dataprep by Trifacta is a serverless service that is well-optimized to prepare and explore data inside the Google Cloud Platform ecosystem. It is an integrated partner service operated by Trifacta and is based on Trifacta Wrangler. Autosuggestion for the next best transformation is available. Cloud Dataflow is used under the covers, which is also a serverless service that can be completely autoscaled with no preprovisioning of clusters and no infrastructure management.

IBM

www.ibm.com

Product name: InfoSphere Advanced Data Preparation

Headquarters: Armonk, New York, U.S.

Geographic presence: North America, EMEA, Asia/Pacific and Latin America

Licensing model: Annual subscription

Number of deployments: Not provided by the vendor

IBM InfoSphere Advanced Data Preparation is used by enterprises that are utilizing multiple data management products in the IBM Infosphere and IBM Cloud Pak for Data product portfolios, are using IBM Watson for their analytics requirements and require the additional capability of data preparation. Jointly developed with Trifacta, the product can be used to prepare and enrich data over data lakes and data warehouses.

Infogix

www.infogix.com

Product name: Data360 Analyze

Headquarters: Naperville, Illinois, U.S.

Geographic presence: North America, EMEA, Asia/Pacific and South America

Licensing model: Sold as desktop and server products:

- Desktop product sold as subscription only, limited to four cores of processing power, and available in two versions (a freemium version and an unlimited, fully supported version).

- Server product sold as both perpetual and subscription, with core-based pricing over and above the four cores, and available in two versions — Automated Server and Enterprise Server.
- No additional charges for data volume or number of data sources, except for the free desktop version that limits data to 2 million rows.

Number of deployments: 600

Infogix provides data preparation capabilities through Data360 Analyze, which includes features from acquired vendor Lavastorm's products. The key differentiator here is that Infogix provides data governance and metadata management solutions as well, through the acquisition of Datum. Key supporting features include security (AD, LDAP integration, role-based access control), data science support (can write code in R or Python) and push down to Apache Spark, Apache Hive, etc. The product is deployable on AWS and GCP.

Informatica

www.informatica.com

Product name: Enterprise Data Preparation (EDP)

Headquarters: Redwood City, California, U.S.

Geographic presence: Asia/Pacific, EMEA, North America, Latin America

Licensing model: Platform cost based on an annual subscription-based license model

Number of deployments: Not provided by the vendor

Informatica EDP is a stand-alone data preparation platform that takes an end-to-end approach to data preparation, i.e., it enables business users to create data preparation flows and IT users (and data engineers) to operationalize self-service models in production environments. The key differentiation for EDP is that it embeds CLAIRE — Informatica's metadata-based ML engine that tracks all forms of metadata including technical, business, performance and social to inform and, in some cases, completely automate repetitive and time-consuming processes. EDP supports data cataloging, data quality, information stewardship (to support governance and policy enforcement) and data security within the data platform, and hence makes operationalization of data models easier in production environments.

Another differentiation is EDP's independent capability, i.e., it can prepare models across any cloud ecosystem and across heterogeneous operational and analytical databases.

Finally, EDP is well-integrated with Informatica's Mass Ingestion and data integration products, which makes it easier for customers to ingest data from multiple sources through Mass Ingestion, then find the data using the catalog functionality and finally prepare the data in their target data store using EDP. Customers typically utilize EDP for allowing SMEs and analysts to explore and wrangle datasets directly in data lake environments to assist with data lake operationalization with intelligent recommendations. EDP can pass its designs and runtime metadata over to any other INFA platform and share the learning information from the metadata. This enables other teams to benefit from the findings of data prepared in EDP.

Lore IO

getlore.io

Product name: Lore IO

Headquarters: Sunnyvale, California, U.S.

Geographic presence: North America, Asia/Pacific, EMEA

Licensing model: Lore IO charges based on the number of unique data sources or schema onboarded for its on-premises and private cloud installations, and charges per terabyte of data stored for its SaaS product. It also offers outcome-based pricing that is linked to the number of output tables or entities.

Number of deployments: 40

Through an AI-optimized workflow, Lore IO's data preparation solution creates a common data model across the enterprise so users can easily find, prepare and deliver data for their data and analytics use cases. Lore IO's target segment typically includes large enterprises who need to operationalize data integration and onboarding. Augmented data preparation is supported through a no-code UI that is enhanced with an auto-recommendation engine that leverages business rules and automatically maps source data so business users can streamline their data preparation efforts. Lore IO helps bootstrap the common data model quickly using a prebuilt model that is optimized for various industries and use cases. User collaboration is supported through the use of data annotations and business rules that automatically create a workflow that allows multiple users to work together. It supports a data catalog for inventorying and indexing data assets and makes them searchable via a semantic search. Lore IO's solution can also perform push-down processing to support in-memory query engines and is available on Hadoop, AWS, GCP and Microsoft Azure.

Modak

www.modak.com

Product name: Modak Nabu

Headquarters: Dubai, UAE

Geographic presence: Asia/Pacific, EMEA, North America

Licensing model: Annual, subscription-based license model charged by the number of processing cores on Spark

Number of deployments: 4

Modak provides data preparation as a capability within a broader data management platform, Nabu, along with other data management capabilities including data cataloging, data integration and data orchestration, among others. Customers can purchase a separate SKU for data preparation that includes data ingestion. Modak is especially proficient at preparing and cataloging data in the life sciences and healthcare sectors. The differentiator here is that Nabu can be used to operationalize data preparation and orchestrate other types of jobs, such as data integration, through support for Spark, StreamSets, Apache NiFi, Informatica Intelligent Cloud Services, AWS Glue, Azure Data Factory and more. Nabu supports bidirectional metadata exchange with leading data catalogs for more robust data preparation and cataloging. Automated data discovery is enabled through DataSpider. Current data governance capabilities include role-based access control, and robust data governance is enabled through integration and strategic partnerships with leading data governance tools. As part of the vendor's roadmap, planned governance features include de-identification for sensitive information.

Quest Software

www.quest.com

Product name: Toad Data Point (TDP)

Headquarters: Aliso Viejo, California, U.S.

Geographic presence: North America, EMEA, Asia/Pacific, Latin America

Licensing model: Sold per "named user" on an annual subscription license with the named user license cost depending on the edition type (base, Pro or Pro+)

Number of deployments: Not provided by the vendor

Quest offers Toad Data Point (TDP) in three editions: the base edition for basic data preparation, the Pro edition, which offers advanced data preparation capabilities (including data transformation, profiling, BI data sources, nonrelational DBMS sources and access to a server edition for teams collaboration) and the Pro+ edition, which also provides access to a library of an additional 170 advanced functions for advanced analytics. Customers buy TDP for its ease of heterogeneous data access and building queries against those sources without coding and without data movement (using federated queries). TDP also allows users to automate and schedule tasks and workflows. TDP offers access to over 50 data sources including relational and nonrelational databases, cloud data stores, Hadoop, enterprise applications (including Salesforce) and BI tools (including SAP BusinessObjects, OBIEE, etc.). An additional differentiator is TDP's capability to assist with data synchronization across different sources to help with data consistency. TDP also supports data profiling capabilities and exporting prepared datasets in various formats including Microsoft Excel, CSV, PDF and HTML, and directly in various analytics tools using APIs. Finally, TDP Pro+ edition includes access to an advanced analytics library that allows users to access advanced mathematical, statistical functions and text algorithms through SQL.

Rapid Insight

www.rapidinsight.com

Product name: Rapid Insight Construct

Headquarters: Conway, New Hampshire, U.S.

Geographic presence: North America

Licensing model:

- $4,000 per user per year for the Rapid Insight Construct tool
- $2,500 per year for the Construct Scheduler Add-on
- $12,000 per year for the Construct Server add-on (see below for details regarding the Scheduler and Server add-ons)

Number of deployments: Not provided by the vendor

Rapid Insight offers a stand-alone data preparation platform, Rapid Insight Construct, to access data from any source and format, integrate and transform it to produce reports, perform ad hoc analysis or create analytics datasets. Construct jobs can be saved, shared and turned into repeatable processes to be run on-demand or on a scheduled basis. Construct has a "preview data" option that is used to preview what the data looks like at each stage of the data preparation process. Construct embeds certain ML algorithms for automation, for example, users can choose a variable that they want to analyze, and Construct will determine which fields have a statistical relationship with that variable and identify the best transformations on that variable. Additionally, metadata can be shared outside the product to convert the data preparation models to ETL processes.

The vendor offers a scheduler (as an add-on) that runs as a service for job scheduling and a server add-on for team collaboration. Finally, Rapid Insight Construct customers can expand to support data visualization and analytics use cases as well, through the Rapid Insight Bridge platform, and can expand to building predictive models through the Predict platform.

SAP

www.sap.com

Product name: SAP Agile Data Preparation, SAP Data Intelligence

Headquarters: Walldorf, Germany

Geographic presence: Asia/Pacific, EMEA, North America, South America

Licensing model: Priced on a memory metric on-premises and capacity units in the cloud (capacity units are a measure of the aggregated memory, compute and storage utilized by the solution)

Number of deployments: Not provided by the vendor

SAP offers SAP Agile Data Preparation as a stand-alone data preparation tool. However, SAP's strategic direction is to provide data preparation capabilities embedded within its broader data management platform, SAP Data Intelligence. SAP Data Intelligence embeds data preparation as one functionality, but it also includes other capabilities, such as data integration, data quality, data governance, data cataloging and embedded ML for automation.

Companies looking to investigate SAP for its data preparation capabilities should note that, while SAP Agile Data Preparation is still available as a stand-alone data preparation tool, SAP Data Intelligence is the solution where future enhancements for data preparation will be made. It is also important to note that SAP has now merged all data preparation capabilities of SAP Agile Data Preparation into SAP Data Intelligence and extended interoperability between the two platforms. This means that datasets prepared in SAP Agile Data Preparation can now be exported to or synchronized with the integrated data in the SAP Data Intelligence. The SAP HANA full use license is a prerequisite for using SAP Agile Data Preparation, but not for using SAP Data Intelligence.

SAP Data Intelligence includes integration capabilities for both SAP and non-SAP data sources. Integration with popular non-SAP sources includes operational databases (including Oracle, Azure SQL Database, IBM Db2, MySQL and Microsoft SQL Server, etc.), cloud data warehouses (including Amazon Redshift and Google's BigQuery, etc.), cloud object stores (including Amazon Simple Storage Service [Amazon S3], Google Cloud Service, HDFC, Azure Data Lake Store, etc.) and other third-party applications including Salesforce, GitHub, Twitter, etc. SAP's data preparation tools are available on-premises and in various public clouds including AWS, GCP, Azure and Alibaba Cloud.

SAS

www.sas.com

Product name: SAS Data Preparation, SAS Data Loader for Hadoop

Headquarters: Cary, North Carolina, U.S.

Geographic presence: North America, EMEA, Asia/Pacific,Latin America

Licensing model:

- SAS Data Preparation is sold on a per-core basis. It can also be licensed per user depending on the deployment style.
- Fees for SAS Data Loader for Hadoop are based on the total number of processing cores in the Hadoop environment. This includes the master, data and edge nodes.

Number of deployments: Not provided by the vendor

SAS Data Loader for Hadoop enables data preparation tasks such as profiling, cleansing, matching, merging and deduplicating directly on Hadoop clusters using MapReduce and Apache Spark, through an intuitive user interface without any specialized coding skills. SAS Data Preparation, on the other hand, works on tables in the SAS Cloud Analytic Services (CAS) in-memory processing environment, where data is loaded into in-memory tables for distributed, parallel execution. Both applications integrate with SAS Data Management, SAS Visual Analytics and SAS Visual Data Mining and Machine Learning. Data quality is embedded into both SAS data preparation tools. SAS Data Preparation also delivers a component called SAS Drive that lets users share artifacts with other users and invites users to work with them in projects. SAS provides metadata bridges that allow SAS object metadata to be consumed in other applications. Data transformation is performed through wizard-driven tasks, while advanced users can also author SAS code for unique data processing tasks. Both these tools can be hosted on AWS and Azure, and on private clouds as well.

Talend

www.talend.com

Product name: Talend Data Preparation

Headquarters: Redwood City, California, U.S.

Geographic presence: Asia/Pacific, EMEA, North America

Licensing model:

- The open-source desktop version is free.
- For the commercial versions, Talend offers a subscription pricing model that is based on a cost per named user license.

Number of deployments: Not provided by the vendor

Talend offers three data preparation options: an open-source desktop version — Talend Data Preparation; and two commercial version options — Talend Cloud Data Preparation (offered as part of the Talend Cloud platform) and Talend Data Preparation (offered as part of Talend Data Fabric). Talend's major differentiation is its rich set of connectivity options, where all connectors are included within the data preparation tool license at no additional cost. Talend Data Preparation utilizes embedded ML algorithms for standardization, cleansing, pattern recognition and reconciliation, and also for offering automated recommendations to guide users through the data preparation process. Talend utilizes Apache Beam as an embedded processing engine, which gives its customers the choice of operationalizing their data preparation flows into any data store (cloud or on-premises) of their choice. Talend's data preparation product is well-integrated with Talend's broader data management platform — Talend Data Fabric (TDF). This allows customers to begin with their self-service data preparation use cases and then operationalize their models through TDF's other data management capabilities including data integration, API management, application integration, information stewardship, data cataloging and data quality. For cloud deployment, Talend also provides a ready-to-run version in all major cloud service provider infrastructures including AWS and Azure.

Tamr

www.tamr.com

Product name: Tamr Enterprise Data Mastering

Headquarters: Cambridge, Massachusetts, U.S.

Geographic presence: Asia/Pacific, EMEA, North America

Licensing model: Tamr typically charges for a three-year license. Pricing is by use cases such as customer mastering, reference data management and others.

Number of deployments: Not provided by the vendor

Tamr Enterprise Data Mastering (TDM) differentiates itself by focusing primarily on enterprise-level agile data mastering and unification use cases. It complements downstream data preparation activity by curating the mastered data being consumed by analysts and data scientists. TDM enables the creation of clean, unified datasets through supervised ML by constructing a probabilistic model that can quickly map, match and categorize data from multiple sources for the purpose of data unification and consolidation. “Mapping” refers to mapping of source attributes to a unified schema. “Matching” refers to record matching, entity resolution, deduplication and mastering. Finally, “categorizing” refers to classifying data using a client-provided taxonomy for downstream analysis.

TDM supports API-based data access for semistructured/unstructured data sources to assist less skilled business users (and SMEs) with unifying and mastering their data at scale (it embeds SPARK as a processing engine). TDM exposes its capabilities as Restful APIs, making it easy to integrate with external data management and analytics tools such as data catalogs, data integration tools, modern analytics and BI tools, and other data preparation tools, as part of the enterprise’s DataOps flow.

For MDM, TDM enables agile data mastering to create golden records of data with minimal coding and using ML to automate the data preparation tasks. TDM allows the data to be left in place, whereas users may share and rate datasets and transformations among themselves, marking records with a degree of confidence to measure the accuracy of the ML model being applied. TDM supports Secure Sockets Layer (SSL)/Secure Shell (SSH)-based encryption, LDAP, SAML and role-based authentication security mechanisms. TDM can be deployed on AWS, Azure and GCP.

Trifacta

www.trifacta.com

Product name: Trifacta Editions include: Free, Standard, Premium and Self-Managed

Headquarters: San Francisco, California, U.S.

Geographic presence: Asia/Pacific, EMEA, North America, South America

Licensing model:

- Trifacta’s Free edition has the same feature capabilities as the Standard edition but limitations on the amount of unique outputs and computation generated through the platform. Trifacta Free is a fully managed SaaS deployment.
- Trifacta’s Standard edition provides the full range of end-user functionality of Trifacta. The licensing model is based on two vectors — unique outputs and computation generated through the platform. Trifacta Standard is a fully managed SaaS deployment.
- Trifacta Premium has all of the end-user functionality of Standard plus advanced security, access controls and connectivity. The pricing model for Premium is consistent with the model for Standard with a slightly higher rate for the additional capabilities. Trifacta Premium is a fully managed SaaS deployment.
- Trifacta Self-Managed enables large enterprises to deploy Trifacta in their private cloud or on-premises environment. It has all of the features of Premium with some customizable capabilities based on the deployment. Pricing is customized to the platform where the product is being deployed and the number of users.

Number of deployments: Over 5,000

Trifacta provides a variety of editions of its data preparation platform supporting various SaaS, private cloud and on-premises computing environments. Its augmented capabilities allow it to recommend data to connect to, infer data structure and schema, recommend joins, define user data access, and automate visualizations for exploration or data quality. A spreadsheet-style grid makes it easy to use, and the ability for analysts to save data preparation steps in a task-based framework makes the process easily repeatable. Data exploration features include sampling, pattern cleaning for data quality assessment and remediation, and automated alerts for data quality issues, among others. Data can be further enriched by instilling ontologies into common data types, thereby improving the overall data preparation process. Data lineage and impact analysis are also delivered as included capabilities for better metadata management. The solution has the option of leaving the data in place and enables push-down processing along with external MPP support.

For security, Trifacta provides support for SSL, SSO, secure impersonation and Kerberos. Trifacta has a strong partner ecosystem and can pull metadata from external data stewardship tools and data catalogs. Trifacta provides out-of-the-box cloud support for AWS, Microsoft Azure and GCP. Trifacta is also widely leveraged by cloud service providers through OEM partnerships to enable data preparation capabilities within their cloud ecosystems, e.g., GCP and IBM have both leveraged Trifacta to support data preparation in their ecosystems. Trifacta is now also available on the AWS Marketplace and supports connectivity and publishing to Amazon Simple Storage Service (Amazon S3) and Amazon Redshift, and deployment on Amazon Elastic Compute Cloud (Amazon EC2).

Market Recommendations

Data and analytics leaders modernizing their data management solution strategy should:

- Assess the roadmaps of their current (or planned) data integration, analytics and BI, data science, and ML platform vendors to determine their roadmap for improving the data preparation capabilities embedded in their product offerings. Many have major product initiatives to supplement their existing products with agile and, in most cases, ML-based data preparation tools that support a range of data types in order to expand their user base and use cases
- Assess the roadmaps of their cloud service providers and their ecosystems for missing functionality or gaps in data cataloging and data preparation and evaluate stand-alone data preparation tools that can successfully plug these gaps.
- Ask their data preparation vendors about their current or planned support for extended data preparation capabilities to improve the interactive, self-service experience and facilitate timely insights. Examples include the inclusion of data science libraries, more-intuitive data preparation workflows, improved governance, collaboration, ML and cataloging.
- Create a process and guidelines for vetting, promoting and operationalizing models developed by business users, and incorporating them into the data integration workflow for data warehousing initiatives (from repeatable and optimized analytics delivery).
- Recognize that data preparation tools will not yet replace the need for robust data integration solutions for all requirements (especially when high volume and complex data transformation and data quality capabilities are needed for the use case).
- Use data preparation tools to bring the necessary agile governance dimension — through data lineage, anomaly detection and cleansing, and shareable metadata — to their enterprise data discovery initiatives. Recognize, however, that changes in roles and responsibilities, processes and user-enablement initiatives will be key to success.

Acronym Key and Glossary Terms

AD	Active Directory
AI	artificial intelligence
AWS	Amazon Web Services
BI	business intelligence
ETL	extraction, transformation and loading
LOB	line of business

ML	machine learning
MDM	master data management
MPP	massively parallel processing
NLP	natural language processing
PaaS	platform as a service
SAML	Security Assertion Markup Language
SSO	single sign-on

Evidence

The findings and vendors included in this research draw on:

- Gartner client inquiry data:
  - During the past four years (from 2016 through 2020, so far), the number of inquiries about data preparation that Gartner has received from clients has been consistently high and has increased by more than 15% year over year. Inquiries have come from a very wide range of industries, and from organizations with varying levels of maturity.
  - We analyzed our inquiry data to profile and analyze vendors that our clients were discussing, as well as curating data from multiple secondary research avenues including Gartner Webinar polling surveys, Gartner client feedback and vendor briefing input.
- Data collected from Gartner's 2019 Data and Analytics Summit inquiries in Dallas, Texas, U.S.; Sydney, Australia; and London, U.K.
- Data collected from interactive vendor briefings conducted for analysts.

Note 1:
Representative Vendor Selection

The vendors and their data preparation products listed in this Market Guide were selected because they offer the key capabilities listed in the Market Description section of this report. They are the vendors about which Gartner has received the most client interest (according to searches on gartner.com and our internal client inquiry service).

Note 2: Gartner's Initial Market Coverage

This Market Guide provides Gartner's initial coverage of the market and focuses on the market definition, rationale for the market and market dynamics.