Data Pipelines
Buyers Guide

Software Provider and Product Assessment

EXECUTIVE SUMMARY

Data Pipelines

The development, testing and deployment of data pipelines is a fundamental accelerator of data-driven strategies. The pipelines enable enterprises to extract data generated by operational applications that run the business and transport it into the analytic data platforms used to analyze the business.

ISG Research defines data pipelines as the systems used to transport, process and deliver data produced by operational data platforms and applications into analytic data platforms and applications for consumption. Healthy data pipelines are necessary to ensure data is ingested, processed and loaded in the sequence required to generate business intelligence and artificial intelligence (AI).

> **Business decisions are becoming increasingly dependent on data-driven processes that require agile, continuous data processing as part of a DataOps approach to data management.**

The concept of the data pipeline is not new. It is, however, increasingly critical as business decisions become more dependent on data-driven processes that require agile, continuous data processing as part of a DataOps approach to data management. In their simplest form, data pipelines move data between production and consumption applications. However, data-driven enterprises are increasingly thinking of the steps involved in extracting, integrating, aggregating, preparing, transforming and loading data as a continual process orchestrated to facilitate data-driven analytics.

The need for more agile data pipelines is driven by the need for real-time data processing. Almost a quarter (22%) of enterprises participating in ISG's Analytics and Data Benchmark Research are currently analyzing data in real time, with an additional 10% analyzing data every hour. More frequent data analysis requires that data is available in a continuous and agile process.

Data pipelines are commonly associated with data integration, which is the software that enables enterprises to extract data from sources such as applications and databases and combine it for analysis to generate business insights. However, data pipelines can move and process data from a single source without any need for integration.

Data integration is performed using data pipelines, but not all data pipelines perform data integration. While the ISG Data Integration Buyers Guide focused specifically on the requirements for data integration pipelines, the Data Pipelines Buyers Guide addresses the wider requirements for data pipelines of all types, as well as higher-level data operations tasks related to the testing and deployment of multiple data pipelines.

Compared to the Data Integration Buyers Guide, the Data Pipelines Buyers Guide places greater emphasis on agile and collaborative practices. This includes integration with the wider ecosystem of DevOps, data management, DataOps and BI and AI tools and applications.

> **The development, testing and deployment of data pipelines can be automated and orchestrated to provide further agility by reducing the need for manual intervention.**

The development, testing and deployment of data pipelines can be automated and orchestrated to provide further agility by reducing the need for manual intervention. Specifically, the batch extraction of data can be scheduled at regular intervals of a set number of minutes or hours, while the various stages in a data pipeline are managed as orchestrated workflows using data engineering workflow management platforms.

Data observability also has a complementary role to play in monitoring the health of data pipelines and associated workflows as well as the quality of the data itself. Many products for data pipeline development, testing and deployment also offer functionality for monitoring and managing pipelines and are integrated with data orchestration and/or observability functionality.

The combination of healthy and well-orchestrated data pipelines and data observability is also complementary to developing and delivering data products, ensuring that data consumers can trust the provenance and quality of data made available across the enterprise.

Traditionally, data pipelines have involved batch extract, transform and load processes designed to extract data from a source (typically a database supporting an operational application), transform it in a dedicated staging area and then load it into a target environment (typically a data warehouse or data lake) for analysis. The need for real-time data processing is driving demand for continuous data processing and more agile data pipelines that are adaptable to changing business conditions and requirements, including the increased reliance on streaming data and events.

There are multiple approaches to increasing the agility of data pipelines. For example, we see an increased focus on extract, load and transform processes that reduce upfront delays in

transforming data by pushing transformation execution to the target data platform. These pipelines involve the more lightweight staging tier, which is required to extract data from the source and load it into the target data platform.

Rather than a separate transformation stage prior to loading, as with an ETL pipeline, ELT pipelines use pushdown optimization, maximizing the data processing functionality and processing power of the target data platform to transform the data. Pushing data transformation execution to the target data platform results in a more agile data extraction and loading phase, which is more adaptable to changing data sources.

> **Pushing data transformation execution to the target data platform results in a more agile data extraction and loading phase.**

Additionally, so-called zero-ETL approaches have emerged to make operational data from a single source available instantly for real-time analytics. Zero-ETL can be seen as a form of ELT that automates extraction and loading and has the potential to remove the need for transformation, assuming that schema is strictly enforced when the data is generated. Meanwhile, reverse ETL tools can help improve actionable responsiveness by extracting transformed and integrated data from the analytic data platforms and loading it back into operational systems.

Both ETL and ELT approaches can be accelerated using change data capture techniques. CDC is similarly not new but has come into greater focus given the increasing need for real-time data processing. As the name suggests, CDC is the process of capturing data changes. Specifically, CDC identifies and tracks changes to tables in the source database as they are inserted, updated or deleted. CDC reduces complexity and increases agility by synchronizing changed data rather than the entire dataset. The data changes can be synchronized incrementally or in a continuous stream.

ISG asserts that by 2026, three-quarters of enterprises will adopt data engineering processes that span data integration, transformation and preparation producing repeatable data pipelines that create more agile information architectures. Additionally, while machine learning (ML) is already used to provide recommendations for building data pipelines, there is also growing interest in applying generative AI to automatically generate or recommend data pipelines in response to natural language explanations of desired outcomes.

There remains a need for traditional batch ETL pipelines, not least to support existing data integration and analytic processes. However, ELT and CDC approaches have a role to play alongside automation and orchestration in increasing data agility, and all enterprises are recommended to explore the potential benefits and evaluate data integration software

providers offering capabilities that support multiple approaches to increase the focus on consumption rather than production-driven data and analytics.

The ISG Buyers Guide™ for Data Pipelines evaluates software providers and products in key areas, including data pipeline development, data pipeline testing, and data pipeline deployment. This research evaluates the following software providers that offer products to address key elements of data pipelines as meet our definition: Airbyte, Alteryx, Astronomer, AWS, BMC, Census, Dagster Labs, Databricks, DataKitchen, DataOps.live, dbt Labs, Google, Hitachi, IBM, Informatica, Infoworks, K2view, Keboola, Mage, Matillion, Microsoft, Nexla, Prefect, Qlik, Rivery, SAP, Y42 and Zoho.

Buyers Guide Overview

For over two decades, ISG Research has conducted market research in a spectrum of areas across business applications, tools and technologies. We have designed the Buyers Guide to provide a balanced perspective of software providers and products that is rooted in an understanding of the business requirements in any enterprise. Utilization of our research methodology and decades of experience enables our Buyers Guide to be an effective method to assess and select software providers and products. The findings of this research undertaking contribute to our comprehensive approach to rating software providers in a manner that is based on the assessments completed by an enterprise.

> **ISG Research has designed the Buyers Guide to provide a balanced perspective of software providers and products that is rooted in an understanding of business requirements in any enterprise.**

The ISG Buyers Guide™ for Data Pipelines is the distillation of over a year of market and product research efforts. It is an assessment of how well software providers' offerings address enterprises' requirements for data pipeline development, testing and deployment software. The index is structured to support a request for information (RFI) that could be used in the request for proposal (RFP) process by incorporating all criteria needed to evaluate, select, utilize and maintain relationships with software providers. An effective product and customer experience with a provider can ensure the best long-term relationship and value achieved from a resource and financial investment.

In this Buyers Guide, ISG Research evaluates the software in seven key categories that are weighted to reflect buyers' needs based on our expertise and research. Five are product-experience related: Adaptability, Capability, Manageability, Reliability, and Usability. In addition, we consider two customer-experience categories: Validation, and Total Cost of Ownership/Return on Investment (TCO/ROI). To assess functionality, one of the components of Capability, we applied the ISG Research Value Index methodology and blueprint, which links the personas and processes for data pipeline development, testing and deployment to an enterprise's requirements.

The structure of the research reflects our understanding that the effective evaluation of software providers and products involves far more than just examining product features, potential revenue or customers generated from a provider's marketing and sales efforts. We believe it is important to take a comprehensive, research-based approach, since making the wrong choice of data pipeline development, testing and deployment technology can raise the total cost of ownership, lower the return on investment and hamper an enterprise's ability to

reach its full performance potential. In addition, this approach can reduce the project's development and deployment time and eliminate the risk of relying on a short list of software providers that does not represent a best fit for your enterprise.

ISG Research believes that an objective review of software providers and products is a critical business strategy for the adoption and implementation of data pipeline development, testing and deployment software and applications. An enterprise's review should include a thorough analysis of both what is possible and what is relevant. We urge enterprises to do a thorough job of evaluating data pipeline development, testing and deployment systems and tools and offer this Buyers Guide as both the results of our in-depth analysis of these providers and as an evaluation methodology.

How To Use This Buyers Guide

Evaluating Software Providers: The Process

We recommend using the Buyers Guide to assess and evaluate new or existing software providers for your enterprise. The market research can be used as an evaluation framework to establish a formal request for information from providers on products and customer experience and will shorten the cycle time when creating an RFI. The steps listed below provide a process that can facilitate best possible outcomes.

1. Define the business case and goals.

   Define the mission and business case for investment and the expected outcomes from your organizational and technological efforts.
2. Specify the business needs.

   Defining the business requirements helps identify what specific capabilities are required with respect to people, processes, information and technology.
3. Assess the required roles and responsibilities.

   Identify the individuals required for success at every level of the enterprise from executives to frontline workers and determine the needs of each.
4. Outline the project's critical path.

   What needs to be done, in what order and who will do it? This outline should make clear the prior dependencies at each step of the project plan.
5. Ascertain the technology approach.

   Determine the business and technology approach that most closely aligns to your enterprise's requirements.
6. Establish software provider evaluation criteria.

   Utilize the product experience: Adaptability, Capability, Manageability, Reliability and Usability, and the customer experience in TCO/ROI and Validation.
7. Evaluate and select the technology properly.

   Weight the categories in the technology evaluation criteria to reflect your enterprise's priorities to determine the short list of software providers and products.
8. Establish the business initiative team to start the project.

   Identify who will lead the project and the members of the team needed to plan and execute it with timelines, priorities and resources.

The Findings

All of the products we evaluated are feature-rich, but not all the capabilities offered by a software provider are equally valuable to types of workers or support everything needed to manage products on a continuous basis. Moreover, the existence of too many capabilities may be a negative factor for an enterprise if it introduces unnecessary complexity. Nonetheless, you may decide that a larger number of features in the product is a plus, especially if some of them match your enterprise's established practices or support an initiative that is driving the purchase of new software.

Factors beyond features and functions or software provider assessments may become a deciding factor. For example, an enterprise may face budget constraints such that the TCO evaluation can tip the balance to one provider or another. This is where the Value Index methodology and the appropriate category weighting can be applied to determine the best fit of software providers and products to your specific needs.

Overall Scoring of Software Providers Across Categories

The research finds Microsoft atop the list, followed by Alteryx and Databricks. Companies that place in the top three of a category earn the designation of Leader. Informatica has done so in five categories; Microsoft in four; Databricks in three; Google and SAP in two; and Alteryx, AWS, DataOps.live, Keboola and Qlik in one category.

The overall representation of the research below places the rating of the Product Experience and Customer Experience on the *x* and *y* axes, respectively, to provide a visual representation and classification of the software providers. Those providers whose Product Experience have a higher weighted performance to the axis in aggregate of the five product categories place farther to the right, while the performance and weighting for the two Customer Experience categories determines placement on the vertical axis. In short, software providers that place closer to the upper-right on this chart performed better than those closer to the lower-left.
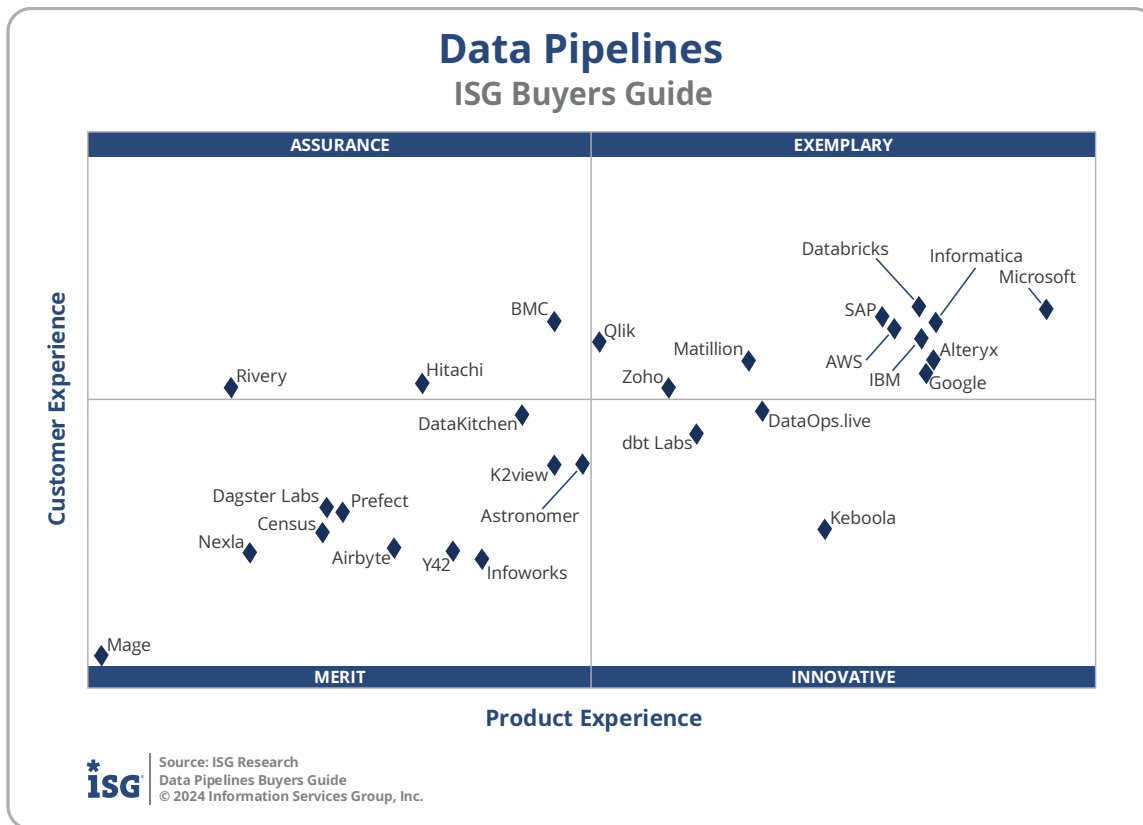
The research places software providers into one of four overall categories: Assurance, Exemplary,

Data Pipelines
Overall

Providers	Grade	Performance
Microsoft	A-	Leader 82.9%
Alteryx	B++	Leader 79.1%
Databricks	B++	Leader 78.5%
IBM	B++	78.0%
Informatica	B++	77.2%
Google	B++	76.9%
AWS	B++	76.7%
SAP	B++	76.6%
DataOps.live	B+	74.1%
Matillion	B+	72.5%
dbt Labs	B+	71.3%
Keboola	B+	70.3%
Zoho	B+	69.8%
BMC	B	68.7%
DataKitchen	B	68.3%
Qlik	B	68.1%
Astronomer	B	65.2%
K2view	B	64.4%
Hitachi	B	64.2%
Y42	B-	61.0%
Infoworks	B-	59.2%
Census	B-	58.0%
Airbyte	B-	57.6%
Dagster Labs	B-	57.4%
Prefect	B-	57.4%
Rivery	B-	57.4%
Nexla	C++	55.8%
Mage	C+	48.8%

Merit or Innovative. This representation classifies providers' overall weighted performance.

Data Pipelines
ISG Buyers Guide

ASSURANCE	EXEMPLARY

Customer Experience / Product Experience

Databricks, Informatica, Microsoft, SAP, Alteryx, AWS, IBM, Google, BMC, Qlik, Matillion, Zoho, Rivery, Hitachi, DataKitchen, DataOps.live, K2view, dbt Labs, Dagster Labs, Prefect, Census, Astronomer, Nexla, Keboola, Airbyte, Y42, Infoworks, Mage

MERIT	INNOVATIVE

Source: ISG Research
Data Pipelines Buyers Guide
© 2024 Information Services Group, Inc.

Exemplary: The categorization and placement of software providers in Exemplary (upper right) represent those that performed the best in meeting the overall Product and Customer Experience requirements. The providers rated Exemplary are: Alteryx, AWS, Databricks, Google, IBM, Informatica, Matillion, Microsoft, Qlik, SAP and Zoho.

Innovative: The categorization and placement of software providers in Innovative (lower right) represent those that performed the best in meeting the overall Product Experience requirements but did not achieve the highest levels of requirements in Customer Experience. The providers rated Innovative are: DataOps.live, dbt Labs and Keboola.

Assurance: The categorization and placement of software providers in Assurance (upper left) represent those that achieved the highest levels in the overall Customer Experience requirements but did not achieve the highest levels of Product Experience. The providers rated Assurance are: BMC, Hitachi and Rivery.

Merit: The categorization of software providers in Merit (lower left) represents those that did not exceed the median of performance in Customer or Product Experience or surpass the threshold for the other three categories. The providers rated Merit are:

Airbyte, Astronomer, Census, Dagster Labs, DataKitchen, Infoworks, K2view, Mage, Nexla, Prefect and Y42.

We warn that close provider placement proximity should not be taken to imply that the packages evaluated are functionally identical or equally well suited for use by every enterprise or for a specific process. Although there is a high degree of commonality in how enterprises handle data pipeline development, testing and deployment, there are many idiosyncrasies and differences in how they do these functions that can make one software provider's offering a better fit than another's for a particular enterprise's needs.

We advise enterprises to assess and evaluate software providers based on organizational requirements and use this research as a supplement to internal evaluation of a provider and products.

Product Experience

The process of researching products to address an enterprise's needs should be comprehensive. Our Value Index methodology examines Product Experience and how it aligns with an enterprise's life cycle of onboarding, configuration, operations, usage and maintenance. Too often, software providers are not evaluated for the entirety of the product; instead, they are evaluated on market execution and vision of the future, which are flawed since they do not represent an enterprise's requirements but how the provider operates. As more software providers orient to a complete product experience, evaluations will be more robust.

The research results in Product Experience are ranked at 80%, or four-fifths, of the overall rating using the specific underlying weighted category performance. Importance was placed on the categories as follows: Usability (10%), Capability (25%), Reliability (15%), Adaptability (15%) and Manageability (15%). This weighting impacted the resulting overall ratings in this research. Microsoft, Informatica, Alteryx and Google were designated Product Experience Leaders.

Data Pipelines
Product Experience

Providers	Grade	Performance
Microsoft	A-	Leader 66.6%
Informatica	B++	Leader 63.4%
Alteryx	B++	Leader 63.2%
Google	B++	Leader 63.2%
IBM	B++	63.0%
Databricks	B++	62.9%
AWS	B++	62.2%
SAP	B++	61.9%
Keboola	B++	60.2%
DataOps.live	B+	58.4%
Matillion	B+	58.0%
dbt Labs	B+	56.5%
Zoho	B+	55.6%
Qlik	B	53.5%
Astronomer	B	53.3%
K2view	B	52.5%
BMC	B	52.3%
DataKitchen	B	51.4%
Infoworks	B	50.2%
Y42	B-	49.4%
Hitachi	B-	48.5%
Airbyte	B-	47.6%
Prefect	B-	46.1%
Dagster Labs	B-	45.8%
Census	B-	45.6%
Nexla	C++	43.4%
Rivery	C++	43.0%
Mage	C+	39.2%

Source: ISG Research
Data Pipelines Buyers Guide
© 2024 Information Services Group, Inc.

Customer Experience

The importance of a customer relationship with a software provider is essential to the actual success of the products and technology. The advancement of the Customer Experience and the entire life cycle an enterprise has with its software provider is critical for ensuring satisfaction in working with that provider. Technology providers that have chief customer officers are more likely to have greater investments in the customer relationship and focus more on their success. These leaders also need to take responsibility for ensuring this commitment is made abundantly clear on the website and in the buying process and customer journey.

The research results in Customer Experience are ranked at 20%, or one-fifth, using the specific underlying weighted category performance as it relates to the framework of commitment and value to the software provider-customer relationship. The two evaluation categories are Validation (10%) and TCO/ROI (10%), which are weighted to represent their importance to the overall research.

The software providers that evaluated the highest overall in the aggregated and weighted Customer Experience categories are Databricks, Microsoft and SAP. These category Leaders best communicate commitment and dedication to customer needs. While not a Leader, Informatica and BMC were also found to meet a broad range of enterprise customer experience requirements.

Data Pipelines
Customer Experience

Providers	Grade	Performance
Databricks	A-	Leader 17.4%
Microsoft	A-	Leader 17.4%
SAP	A-	Leader 17.2%
Informatica	A-	17.1%
BMC	A-	17.1%
AWS	A-	16.9%
IBM	A-	16.7%
Qlik	A-	16.6%
Matillion	B++	16.2%
Alteryx	B++	16.1%
Google	B++	16.0%
Hitachi	B++	15.5%
Rivery	B++	15.2%
Zoho	B++	15.2%
DataOps.live	B++	15.2%
DataKitchen	B+	14.9%
dbt Labs	B+	14.4%
Astronomer	B	13.7%
K2view	B	13.7%
Dagster Labs	B	12.6%
Prefect	B	12.5%
Keboola	B-	12.2%
Census	B-	12.0%
Airbyte	B-	11.7%
Y42	B-	11.6%
Nexla	B-	11.5%
Infoworks	B-	11.4%
Mage	C+	8.9%

Source: ISG Research
Data Pipelines Buyers Guide
© 2024 Information Services Group, Inc.

Software providers that did not perform well in this category were unable to provide sufficient customer case studies to demonstrate success or articulate their commitment to customer experience and an enterprise's journey. The selection of a software provider means a continuous investment by the enterprise, so a holistic evaluation must include examination of how they support their customer experience.

Appendix: Software Provider Inclusion

For inclusion in the ISG Buyers Guide™ for Data Pipelines in 2024, a software provider must be in good standing financially and ethically, have at least $10 million in annual or projected revenue verified using independent sources, sell products and provide support on at least two continents, and have at least 50 employees. The principal source of the relevant business unit's revenue must be software-related, and there must have been at least one major software release in the last 18 months.

The software provider must provide a product or products that support agile and collaborative data operations and are marketed as addressing at least one of the following functional areas, which are mapped into Buyers Guide capability criteria: data pipeline development, data pipeline testing and data pipeline management.

The development, testing and deployment of data pipelines enables enterprises to extract data from the operational applications and data platforms designed to run the business and load, integrate and transform it into the analytic data platforms and tools used to analyze the business. Data pipelines are a fundamental accelerator of data-driven strategies, and today's analytics environments require agile data pipelines that can traverse multiple data-processing locations and evolve with business needs.

To be included in this Buyers Guide requires functionality that addresses the following sections of the capabilities document:
- Data pipeline development
- Data pipeline testing
- Data pipeline deployment

The research is designed to be independent of the specifics of software provider packaging and pricing. To represent the real-world environment in which businesses operate, we include providers that offer suites or packages of products that may include relevant individual modules or applications. If a software provider is actively marketing, selling and developing a product for the general market and it is reflected on the provider's website that the product is within the scope of the research, that provider is automatically evaluated for inclusion.

All software providers that offer relevant data pipeline products and meet the inclusion requirements were invited to participate in the evaluation process at no cost to them.

Software providers that meet our inclusion criteria but did not completely participate in our Buyers Guide were assessed solely on publicly available information. As this could have a significant impact on classification and ratings, we recommend additional scrutiny when evaluating those providers.

Products Evaluated

Provider	Product Names	Version	Release Month/Year
Airbyte	Airbyte Cloud	1.10	September 2024
Alteryx	Analytics Cloud	N/A	October 2024
Astronomer	Astro	N/A	October 2024
AWS	AWS Glue Amazon Managed Workflows for Apache Airflow	N/A N/A	September 2024 September 2024
BMC	Control-M	9.0.21.300	October 2024
Census	Census	N/A	September 2024
Dagster Labs	Dagster+	1.8.12	October 2024
Databricks	Data Intelligence Platform	N/A	October 2024
DataKitchen	DataOps TestGen	2.15.3	October 2024
DataOps.live	DataOps.live	October 2024	October 2024
dbt Labs	dbt	October 2024	October 2024
Google	Cloud Data Fusion Cloud Dataflow	N/A N/A	October 2024 September 2024
Hitachi	Pentaho Data Integration	10.2	September 2024
IBM	Cloud Pak for Data	5.0	September 2024
Informatica	Intelligent Data Management Cloud	October 2024	October 2024
Infoworks	Infoworks	6.1.0	September 2024
K2view	Data Product Platform	8.1.1	October 2024
Keboola	Keboola	N/A	November 2024

Mage	Mage	0.9.72	June 2024
Matillion	Data Productivity Cloud	N/A	October 2024
Microsoft	Fabric	October 2024	October 2024
Nexla	Nexla	N/A	October 2024
Prefect	Prefect Cloud	3.0	September 2024
Qlik	Talend Cloud	N/A	October 2024
Rivery	Rivery	October 2024	October 2024
SAP	Data Intelligence Cloud, Datasphere	N/A 2024.20	April 2024, September 2024
Y42	Y42	N/A	July 2024
Zoho	DataPrep	2.0	October 2024

Providers of Promise

We did not include software providers that, as a result of our research and analysis, did not satisfy the criteria for inclusion in this Buyers Guide. These are listed below as "Providers of Promise."

Provider	Product	Annual Revenue >$10M	Operates on 2 Continents	At Least 50 Employees	GA Product/ Documentation
Arch Data	Meltano	No	Yes	No	Yes
Ascend	Data Automation Cloud	No	Yes	No	Yes
Datacoves	Datacoves	No	Yes	No	Yes
Datafold	Datafold	No	Yes	No	Yes
FirstEigen	DataBuck	No	Yes	No	Yes
Integrate.io	Data Observability	No	Yes	No	Yes
Kleene	Kleene	No	Yes	No	Yes
Mozart Data	Mozart Data	No	Yes	No	Yes
Pipedream	Pipedream	No	Yes	No	Yes
Promethium	Promethium	No	Yes	No	Yes
PurpleCube AI	PurpleCube AI	No	Yes	No	Yes
RightData	DataFactory	Yes	Yes	Yes	No
Saturam	Qualdo, Piperr	No	Yes	Yes	No
Switchboard Software	Data Automation	No	Yes	No	Yes
Torana	iceDQ	No	Yes	Yes	No
Upsolver	Upsolver	No	Yes	No	Yes

About ISG Software Research

ISG Software Research provides expert market insights on vertical industries, business, AI and IT through comprehensive consulting, advisory and research services with world-class industry analysts and client experience. Our ISG Buyers Guides offer comprehensive ratings and insights into technology providers and products. Explore our research at www.isg-research.net.

About ISG Research

ISG Research provides subscription research, advisory consulting and executive event services focused on market trends and disruptive technologies driving change in business computing. ISG Research delivers guidance that helps businesses accelerate growth and create more value. For more information about ISG Research subscriptions, please email contact@isg-one.com.

About ISG

ISG (Information Services Group) (Nasdaq: III) is a leading global technology research and advisory firm. A trusted business partner to more than 900 clients, including more than 75 of the world's top 100 enterprises, ISG is committed to helping corporations, public sector organizations, and service and technology providers achieve operational excellence and faster growth. The firm specializes in digital transformation services, including AI and automation, cloud and data analytics; sourcing advisory; managed governance and risk services; network carrier services; strategy and operations design; change management; market intelligence and technology research and analysis. Founded in 2006 and based in Stamford, Conn., ISG employs 1,600 digital-ready professionals operating in more than 20 countries—a global team known for its innovative thinking, market influence, deep industry and technology expertise, and world-class research and analytical capabilities based on the industry's most comprehensive marketplace data. For more information, visit isg-one.com.