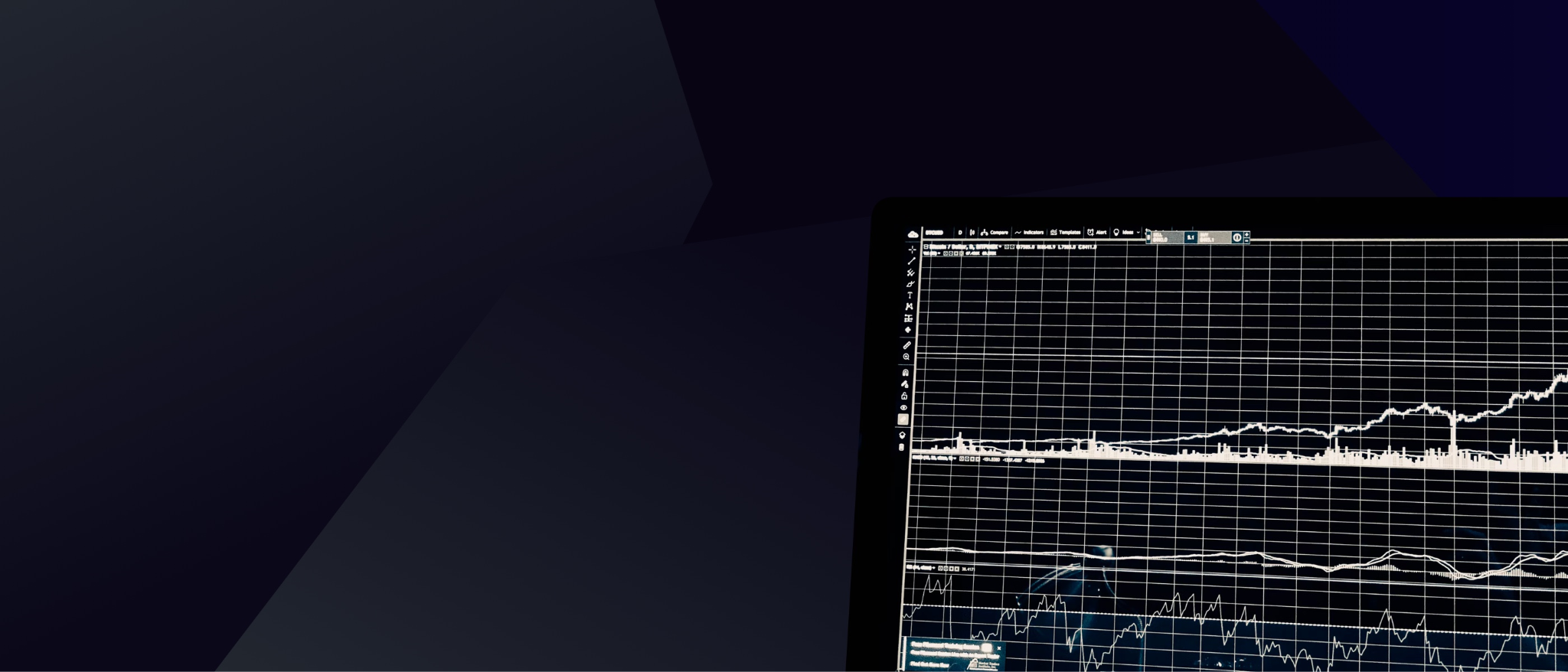
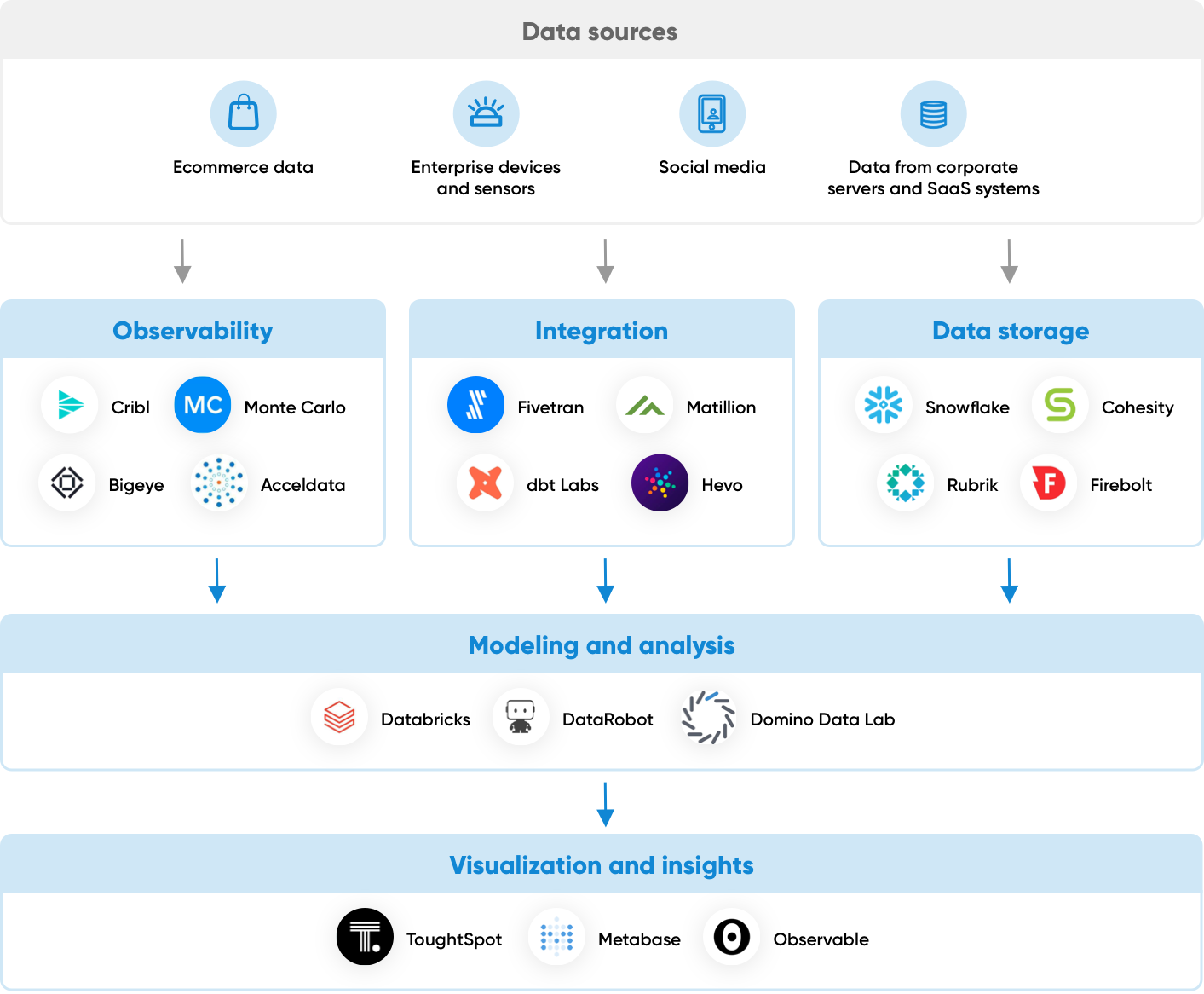
# **Data Infrastructure and Analytics: Overview**



## **Data Infrastructure: The backbone of contemporary data-driven enterprises**

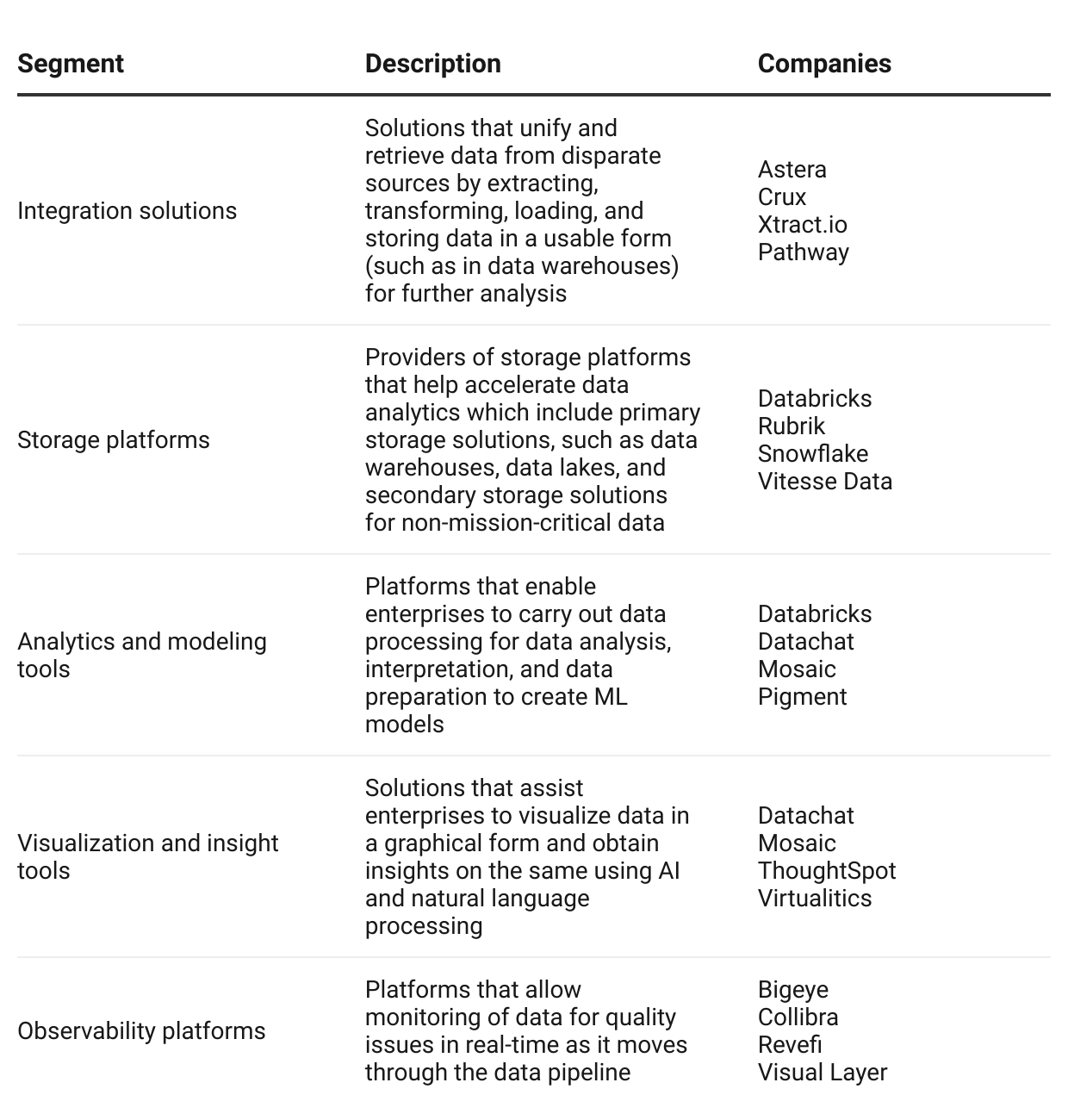
Data is ubiquitous, comes in many forms, and its significance cannot be overstated. It underpins every facet of business decision-making, from the neighborhood convenience store deciding when to restock its best-selling items to Silicon Valley tech companies preparing for the next big innovation. Enterprises collect data from multiple sources, including point-of-sale systems, cloud environments, social media accounts, and IoT devices. Moreover, customers generate petabytes of data every day and companies harness this information to uncover new opportunities, boost revenue, and improve efficiencies to gain a competitive edge over their rivals.

This industry hub focuses primarily on the “infrastructure” or “data stack” that encompasses the intricate network of tools and solutions helping organizations and data teams assess and remediate data quality and properly structure the data for secure storage, aggregation, and distribution for analytics, reporting, visualizations, and insight generation that influence informed decision making. These also help enterprises curate data to build and train their own machine-learning (ML) models.



## **Key segments of Data Infrastructure**

Companies offering solutions and tools in the Data Infrastructure space either target individual processes or provide integrated platforms that cater to multiple processes. These solution providers assist enterprises in acquiring data from diverse sources, securely storing it for subsequent analysis and extracting valuable insights. This category also encompasses companies that offer solutions for monitoring and curating data, thus enhancing its quality and enabling data scientists and ML practitioners to produce higher-quality ML models and insights.



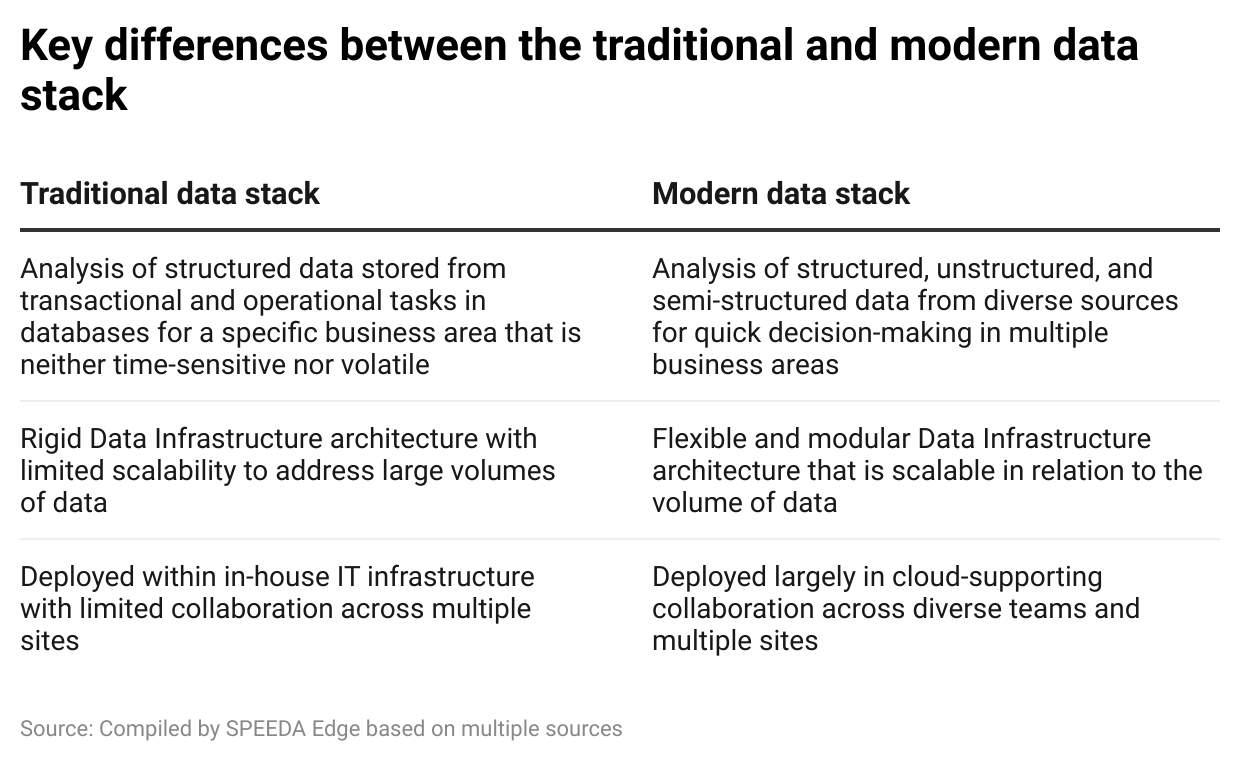
*For more information on tools that help organizations build, train, and deploy ML and large language models (LLMs) for GenAI applications, please visit the* [Machine Learning Infrastructure](https://sp-edge.com/industry/160) *and* [Generative AI Infrastructure](https://sp-edge.com/industry/178) *industry hubs, respectively.*

## 

## **The legacy data stack is falling behind today’s pace of innovation**

Traditional approaches to data analysis have numerous limitations. It often begins with the manual compilation of data from various sources and requires the identification of duplicates and inconsistencies. Moreover, traditional data stacks that depended on a company's in-house data centers lacked the flexibility and scalability needed for collaboration across various locations. Furthermore, traditional data analysis tools were designed primarily for structured data, which is inadequate for modern-day organizations, as nearly 80% of all enterprise data is unstructured in nature.

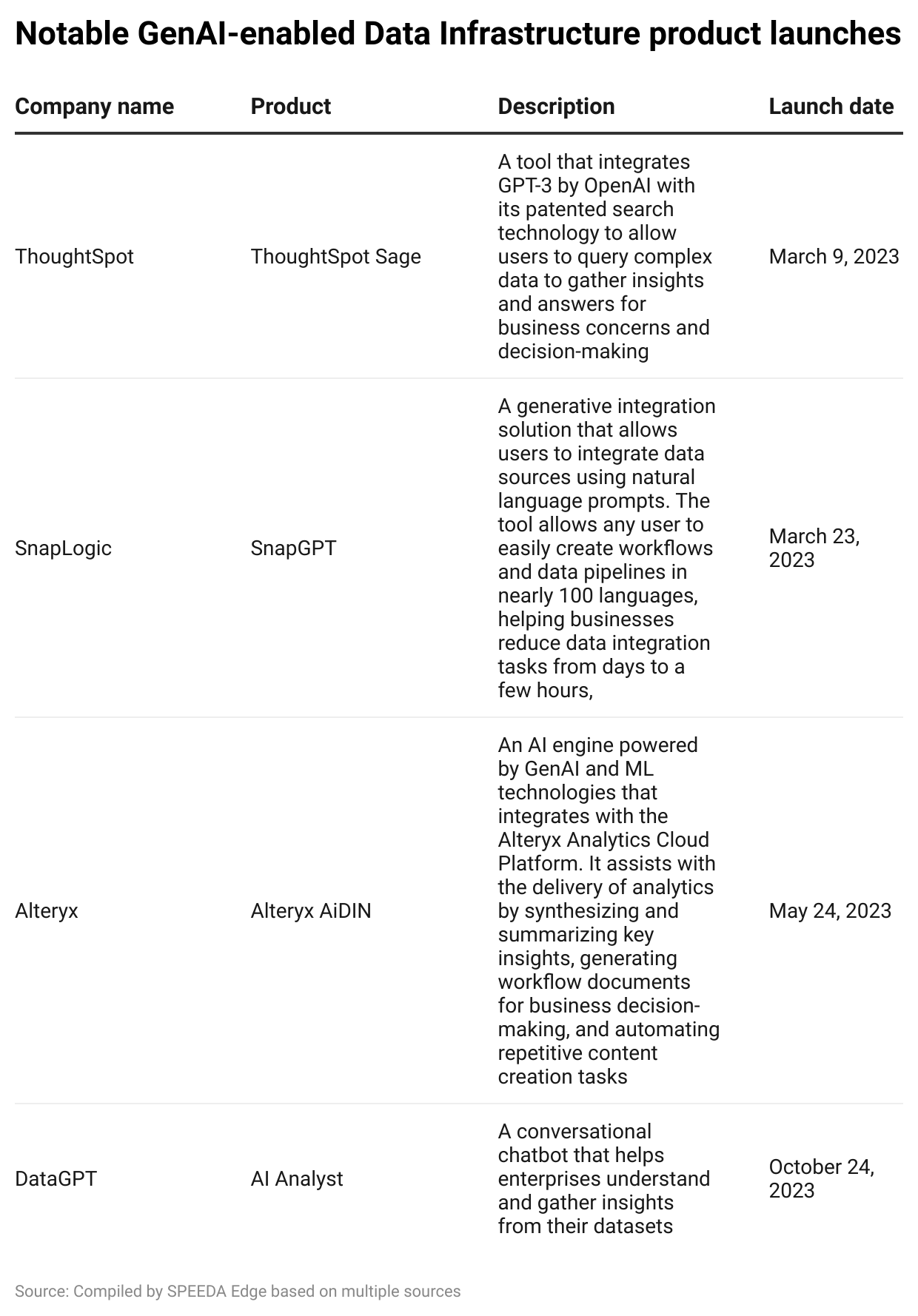
Notably, organizations were frequently locked into long-term contracts with specific vendors since their solutions were non-modular and created with a defined set of shared business objectives in mind. This hampered the full use of data for decision-making, leaving little to no room to unearth hidden insights within the data.

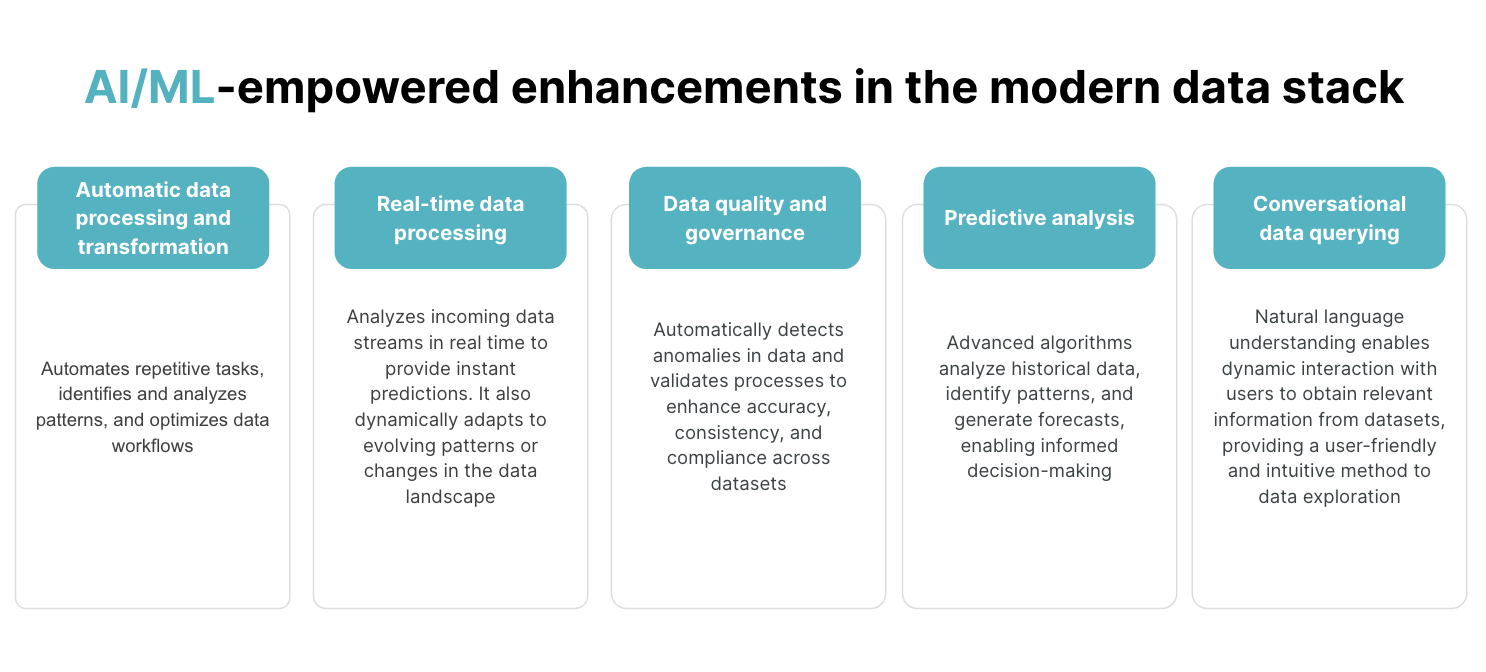


## **AI and ML are influential in transforming Data Infrastructure stack**

Advancements in AI and ML algorithms have played a significant role in the development and evolution of the modern data stack. It has enabled Data Infrastructure platforms to automate data processing and transformation processes, along with data quality and governance operations. It also enabled these platforms to process data in real-time with predictive analysis capabilities—previously impossible with traditional data platforms—providing organizations with insights and enabling them to make swift decisions in response to the evolving competitive landscape.

Moreover, the recent development of GenAI has led many solution providers to explore ways to either develop their own GenAI features or collaborate with companies specializing in GenAI to better their data analysis solutions. These features allow users regardless of their skills and knowledge to query the data in a conversational language to obtain contextualized explanations for data inputs and outputs, as well as obtain insights and identify trends, anomalies, and outliers present within the data.





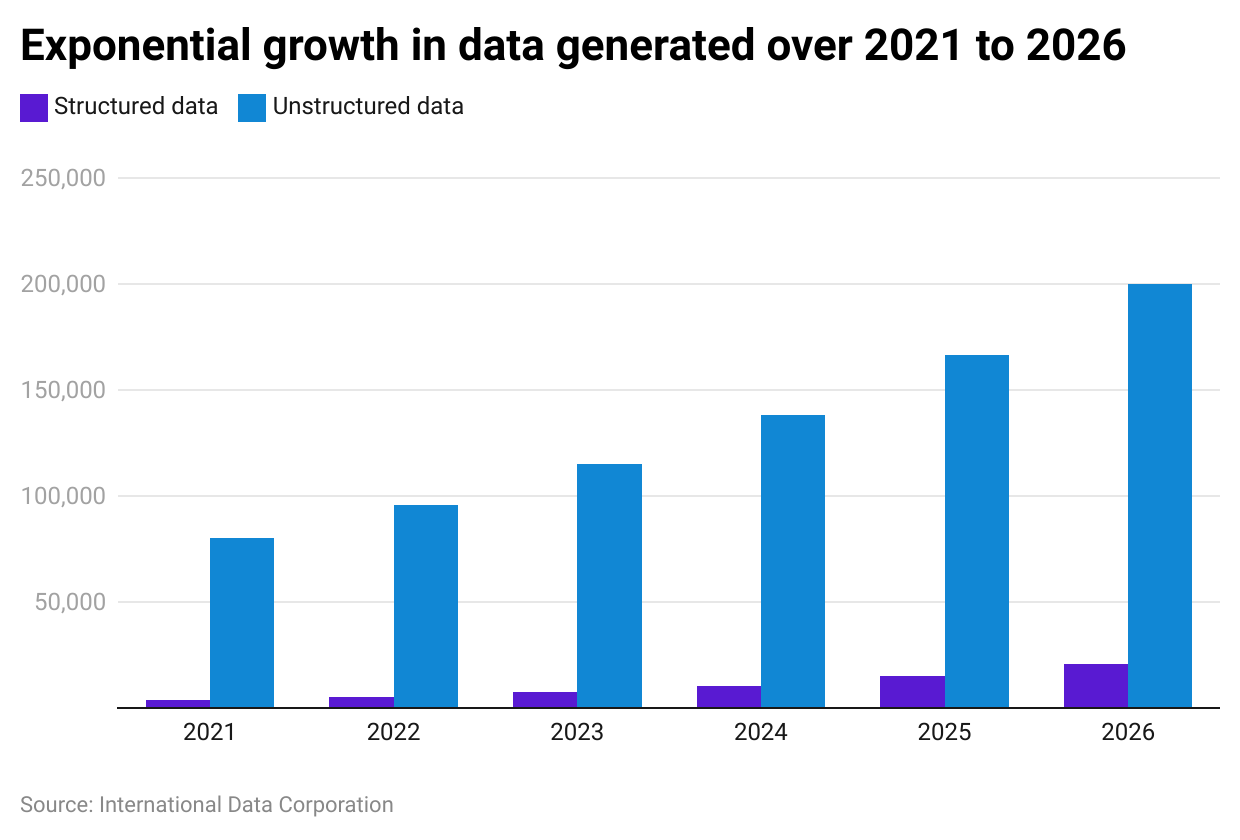
Source: Compiled by SPEEDA Edge based on multiple sources

# **Driving factors**

## **1. Exponential growth in data volumes spurs demand for Data Infrastructure solutions**

In today's digital era, humans actively and passively generate vast amounts of data. According to the International Data Corporation, roughly 84,000 exabytes of data were generated in 2021, and this figure is projected to soar by over 2.5x, reaching 221,000 exabytes by 2026; approximately 90% of this data is unstructured.

The challenge is not only the sheer volume of data that needs to be processed but also the complexity of integrating it into the data pipelines and workloads of enterprises. In addition, vetting and validating this data before analysis and other essential tasks, such as constructing AI and ML models, also pose significant challenges. Notably, one-fifth of large enterprises report extracting data from over 1,000 sources, with an average of 400 data sources, while data teams usually dedicate around 20% of their time on data acquisition and preparation and another 40% in critical tasks of data verification and validation.



Data Infrastructure solution providers step in to address these pain points, offering solutions to enterprises regardless of their size. These solutions help in retrieving and integrating data from various sources, breaking down data silos, and providing scalable cloud platforms to handle heavy data workloads. Moreover, these platforms also harness AI and ML to automatically transform data by detecting duplicates and anomalies and validating them. These technological advances not only save enterprises time and money but also offer improved visibility into the quality of their data. For example, IBM Cloud Pak for Data, an insight platform that combines data management with data science and AI development, is estimated to reduce infrastructure management effort by around 65% to 85% and extract, transform, and load (ETL) requests by 25% to 65%.

## **2. The success of AI and ML applications—and more recently GenAI—relies on high-quality data**

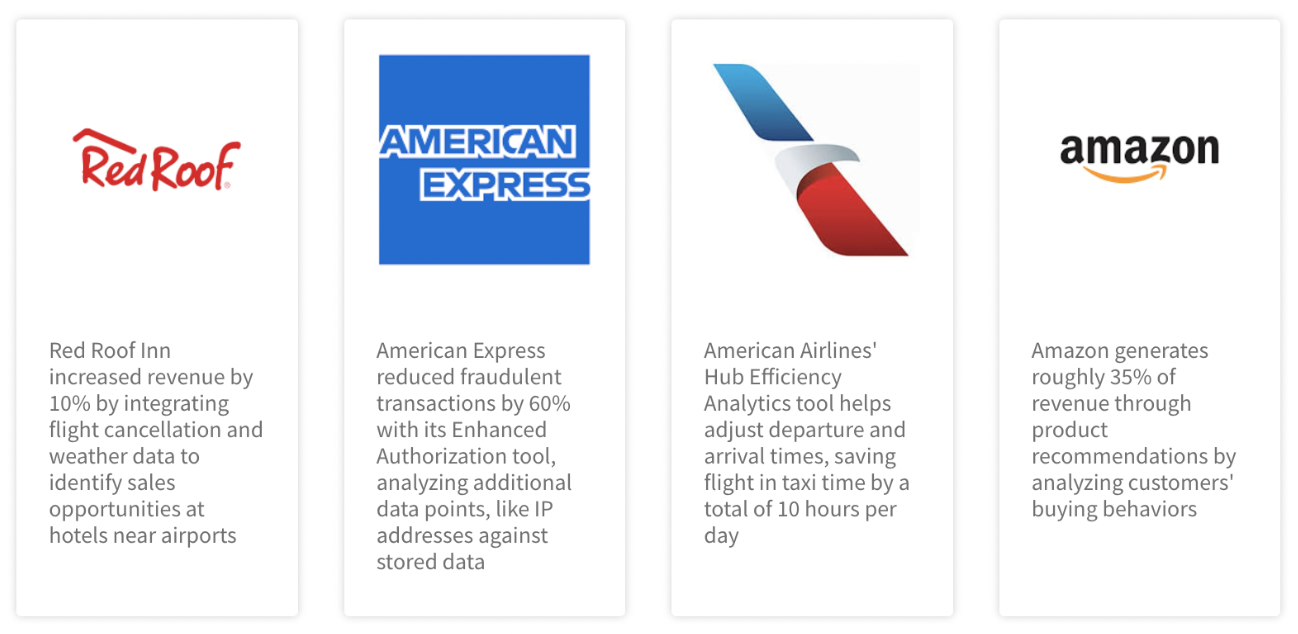
Demand for ML and AI applications has skyrocketed in recent years due to their transformative potential across industries, enhancing customer experiences to optimize operations and making data-driven decisions. The effectiveness of these applications is directly linked to the volume and quality of data used to train them as they discover patterns and draw inferences based on them. A prime example is GPT-3.5, a [large language model (LLM)](https://sp-edge.com/insights/16079) developed by [OpenAI](https://sp-edge.com/companies/362567), which was trained on 570 GB of data, encompassing nearly 300 billion words.

According to [McKinsey](https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-data-dividend-fueling-generative-ai), 72% of chief data officers at leading organizations have pinpointed data management as a significant hurdle in scaling AI use cases. Additionally, the cost of training AI and ML models can escalate rapidly, underscoring the need for accurate, diverse, and vetted data. The cost of training foundation models is estimated to be north of USD 100 million as of September 2023, while the cost of training the next generation of LLMs is expected to surpass USD 1 billion. As a result, the development and deployment of ML and AI systems is dependent on robust data infrastructure platforms, which serve as the foundation for the integration and curation of the massive datasets needed for training and continuing operations of AI and ML applications.

*For more information on companies providing synthetic data and data annotation services and platforms for building AI and ML models, please visit the* [Machine Learning Infrastructure](https://sp-edge.com/industry/160)*. For details on platforms that assist in building LLMs and GenAI applications, please visit the* [Generative AI Infrastructure](https://sp-edge.com/industry/178) *industry hub.*

## **3. Data-driven decision-making is imperative to gain a competitive edge**

Enterprises are increasingly using data to inform and shape every element of decision-making. They employ data analytics to derive important insights from the data they collect, gaining knowledge about customer behavior, market trends, and internal operations. Furthermore, data-driven analysis improves operational efficiencies by finding areas for improvement, streamlining procedures, and lowering costs. The results of a data-driven approach are tangible, as companies are 6x more likely to retain clients and have reported an 8% increase in profits.



Source: Compiled by SPEEDA Edge based on multiple sources

To support these initiatives, enterprises heavily depend on Data Infrastructure platforms for storing data and obtaining insights. Furthermore, these platforms will maintain a crucial role in the future as data analytics extends beyond data teams to individual employees. This shift is due to advancements in GenAI tools for data analysis, which enable non-technical users to query data for insights in natural language.

# **Risks to growth**

## **1. Evolving data and privacy regulations could limit the use of data**

Evolving data and privacy regulations are increasingly shaping the way organizations can use and handle data, potentially imposing limitations on data utilization. In the US, regulations such as the California Consumer Privacy Act (CCPA) and the California Privacy Rights Act (CPRA) grant consumers greater control over their personal data, allowing them to request the deletion of data and choose to opt out of data sales. Meanwhile, in Europe, the General Data Protection Regulation (GDPR) imposes strict requirements on data handling, requiring explicit consent for data processing, mandating data breach notifications, and enabling individuals to access and delete their data.

Further regulations at a national and international level could limit the features and use of Data Infrastructure platforms by enterprises while imposing additional costs on regulatory compliance requirements.

*For more information on data governance platforms and the use of privacy-preserving tools to process and collaborate on data in a compliant manner, please visit the* [Digital Privacy Tools](https://sp-edge.com/industry/34) *industry hub.*

## **2. Tool sprawl could lead to fragmented data pipelines and high operational costs**

A notable criticism of today’s data stack revolves around the overwhelming abundance of tools available to organizations, making the selection of the right combination a daunting task. The use of disparate tools by different teams can lead to data silos, operational inefficiencies, compromised data quality, and security vulnerabilities due to overlapping functionality and poor integration. This also adds complexity to the procurement process, involving multiple vendor negotiations, licenses, and subscription management. Moreover, the cumulative costs of licenses, infrastructure, training, support, maintenance, and other operational expenses can add up quickly, creating financial challenges for organizations.

*Last updated: November 2023*

©2024 Uzabase, Inc. All Rights Reserved. The information contained herein: (1) is proprietary to Uzabase Inc. and/or its content providers; (2) may not be copied or distributed; and (3) is not warranted to be accurate, complete or timely. Neither Uzabase Inc. nor its content providers are responsible for any damages or losses arising from any use of this information.