



# The Data Integration Advantage: Building a Foundation for Scalable AI

Authored by

*Girish Pancha, Co-Founder, Chief Executive Officer*

*Arvind Prabhakar, Co-Founder, Chief Product Officer*

[www.streamsets.com](http://www.streamsets.com)

## Introduction

For the last decade or so, data has been the business world's darling. Curious why your customers are unhappy? Look at the data. Wondering what your next market should be? The data will tell you. Want to find out who your best-performing employees are? You know what to do.

Now, there's a (not so) new kid in town that's dominating the conversation: AI. Generative AI has ignited imaginations across the world. As the first widely available application that lets anyone talk to an AI about anything — and get coherent, even clever answers — AI has moved from the abstract to an everyday reality.

But while AI may be overtaking public discourse, data is (of course) not going anywhere. That's because the success of AI projects is not simply a result of innovative algorithms or machine learning models; it fundamentally relies on mass quantities of accessible, reliable data. AI, ML, and analytics output are meaningful only if the data they operate on is valid and observable across the whole lifecycle — sample data for exploration, test and training data for experimentation, and production data for evaluation.

As AI initiatives become more ambitious and scale across organizations, the demand for connected, quality, governed data increases in parallel. Modern data integration is the critical backbone for successfully scaling AI. And with 72% of Fortune 500 business leaders planning to incorporate generative AI within the next three years<sup>1</sup>, it's time to get data integration right.

In this piece, we'll explore:

- **The state of AI in the enterprise**
- **Challenges of scaling AI**
- **How modern data integration can remove AI scaling challenges**
- **Moving beyond data integration for even better AI results**

Read on to learn about data integration's vital role in the quest to scale AI.

**72% of technology executives say that should their companies fail to achieve their AI goals, data issues are more likely than not to be the reason.**

[CIO Vision 2025: Bridging the Gap Between BI and AI, MIT Technology Review Insights](#)

<sup>1</sup> [Beyond Hypotheticals: Understanding the Real Possibilities of Generative AI, Insight](#)

## The State of AI in the Enterprise


For years, enterprises have been using AI in pockets around the enterprise. It's made great strides in:

- Improving customer experience through chatbots and virtual assistants powered by natural language processing (NLP) that provide instant, personalized customer service 24x7.
- Optimizing supply chain processes by predicting demand, optimizing delivery routes, and identifying potential disruptions.
- Identifying when machinery is likely to fail (predictive maintenance) to carry out maintenance before a breakdown occurs.
- Expediting research and development processes, reducing the time to market for products and services.
- Detecting fraud, evaluating credit risk, and anticipating market changes with machine learning algorithms that identify patterns in historical data.

However, most enterprise AI usage is limited to very specific use cases and departments. BCG

### Maturity of AI uses cases across industries

| Functional categories                                     | Consumer | Energy | Financial institutions | Health care | Industrial goods | Insurance | Public sector | Tech | Telco |
|---|----------|--------|------------------------|-------------|------------------|-----------|---------------|------|-------|
| Supply chain and network (e.g., inventory optimization)   | 42       | 39     | 40                     | 41          | 39               | 38        | 39            | 36   | 37    |
| Enterprise (e.g., HR analytics)                           | 43       | 38     | 37                     | 41          | 40               | 38        | 39            | 36   | 36    |
| Manufacturing (e.g., predictive maintenance)              | 40       | 37     | 38                     | 37          | 39               | 37        | 36            | 37   | 42    |
| Marketing and customer experience (e.g., personalization) | 40       | 38     | 38                     | 37          | 38               | 37        | 37            | 38   | 38    |
| Products and offers (e.g., pricing)                       | 42       | 39     | 35                     | 38          | 37               | 36        | 36            | 38   | 39    |
| Risk (e.g., fraud detection)                              | 45       | 39     | 40                     | 41          | 39               | 40        | 37            | 37   | 37    |
| Overall   | 42       | 37     | 38                     | 38          | 39               | 37        | 36            | 37   | 37    |

0  100

Source: BCG Digital Acceleration Index global study, 2022

found that only 11% of companies have realized significant value from AI initiatives, and most have failed to scale AI beyond pilots.<sup>2</sup>

Their 2022 digital acceleration index — a survey of 2700 companies — paints a picture of AI initiatives stuck in the early stages.<sup>3</sup>

However, there were 'leaders' in scaling and generating AI value among this group. BCG found that one of the primary characteristics of those leaders was making "data and technology accessible across the organization, avoiding siloed and incompatible tech stacks and standalone databases that impede scaling."

**78% of enterprise technology leaders said that scaling AI and machine learning use cases to create business value is the top priority of their enterprise data strategy over the next three years<sup>4</sup>**

CIO Vision 2025:  
Bridging the Gap Between BI and AI,  
MIT Technology Review Insights

<sup>2</sup> Artificial Intelligence, Ready to Ride the Wave?, BCG

<sup>3</sup> Scaling AI Pays Off, No Matter the Investment, BCG

<sup>4</sup> CIO Vision 2025: Bridging the Gap Between BI and AI, MIT Technology Review Insights.

## The Challenges of Scaling AI

While there are many challenges in scaling AI — cost, lack of talent, trust and ethics — data quality and availability are arguably the biggest hurdles. In fact, 72% of technology executives surveyed in a recent MIT study say that should their companies fail to achieve their AI goals, data issues are more likely than not to be the reason,<sup>5</sup> and 61% of respondents in an IBM survey said their data is not ready for AI.<sup>6</sup>

AI models rely on a constant influx of high-quality data for training and inference. But, organizations often grapple with data quality issues such as incomplete and inaccurate data. Another problem is integrating relevant data from different sources across the organization, such as mainframes, customer relationship management (CRM) systems, enterprise data warehouses and data lakes, business intelligence platforms, external systems, third-party data, and more.

To make matters even more complex, AI/ML models are not static; they require ongoing monitoring and maintenance to ensure performance and reliability. Monitoring for concept drift, model decay, and performance degradation is essential. Regular updates and retraining may be necessary to adapt models to evolving data patterns or changes in the operational environment. As such, organizations must establish processes to manage version

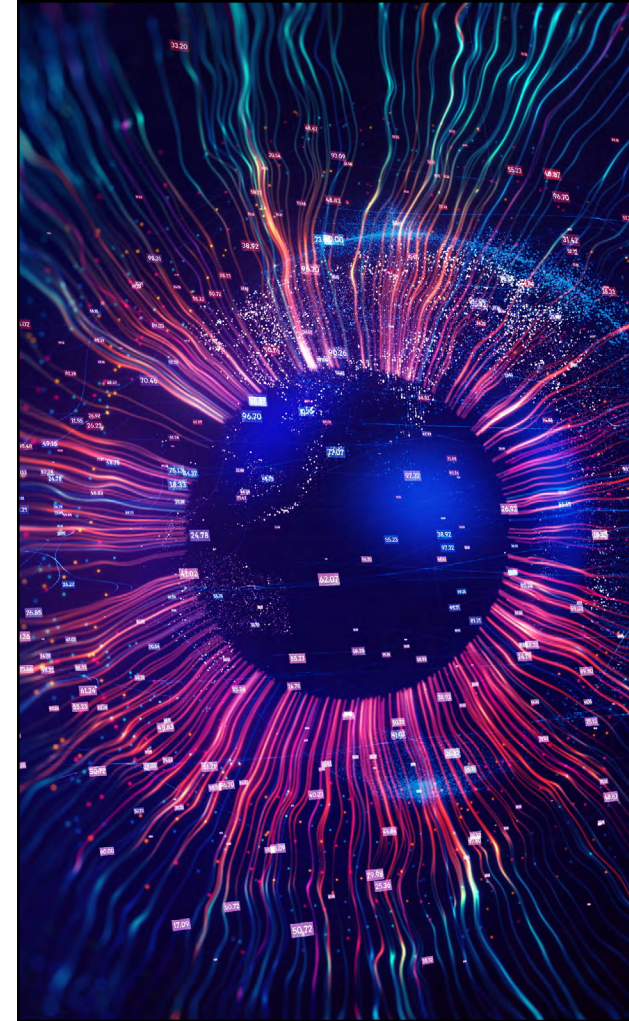
control, model updates, and performance tracking.

Today, most organizations handle these processes manually. They create manual workflows around retraining data, use new datasets, identify boundary conditions or fringe predictions that don't match the norm, and then make the best guess as to the right time to retrain the model. Clearly, this is an imprecise science that can lead to subpar outcomes.

Given these challenges, a solid data foundation is essential for AI/ML models to function properly over the long term. The ability to easily access and share high-quality data — real-time or batch — across the organization securely is essential for building an AI-powered application that's relevant, accurate, and scalable

<sup>5</sup> [CIO Vision 2025: Bridging the Gap Between BI and AI.](#)  
[MIT Technology Review Insights](#)

<sup>6</sup> [AI in the Enterprise, IBM.](#)



# How Modern Data Integration Solutions Can Remove AI Scaling Challenges

A recent PWC survey<sup>7</sup> found that the top tech-related challenge for AI is identifying, collecting, or aggregating data from across the company, ensuring its completeness and accuracy in preparation for use in AI. This was followed closely by making sure all data in AI systems meets regulatory requirements for privacy and data protection and integrating AI and analytics systems to gain business insights.

As you upgrade your technology and architecture, they suggest focusing on two imperatives: integration and data. “With technology tools that help you overcome your data challenges, you can achieve much faster (and much more cost-effective) operationalizing of AI.”

Let’s look at how data integration technology can help with challenges specific to scaling AI/ML.

<sup>7</sup> [To operationalize AI, reorganize in these three ways.](#)  
PWC

|  | AI Scaling Challenge  | How Modern Data Integration Helps   |
|--|---|---|
| <b>Data silos</b>                                    | Data gets trapped in departmental silos, legacy systems, and cloud apps in varying formats. This data fragmentation makes it hard to aggregate the large, diverse datasets needed to train accurate AI models.  | A modern data integration solution will provide connectors to gather data from various data stores and infrastructure, including legacy systems like mainframes. It can then transform disparate data formats into a consistent, analysis-ready format.   |
| <b>Data quality and availability</b>                 | AI systems rely heavily on vast amounts of high-quality and relevant data for training and making accurate predictions. Data often has issues like missing fields, outliers, duplicates, inconsistencies, and lack of context. Low-quality data leads to poor model performance.  | With data integration, businesses can automate data cleansing tasks like handling nulls, deduplication, normalization, and validation. Cleaning the data used for AI training and decision-making reduces the risk of biased or inaccurate models.  |
| <b>Data security &amp; privacy</b>                   | Training data may contain personal and sensitive information requiring protections like encryption, anonymization, and access control.  | Data integration tools can secure data movement with encryption and anonymize data by masking fields. They should be compatible with data access and LDAP tools for extra security.   |
| <b>Data context</b>                                  | AI models rely on metadata like data definitions, datatypes, hierarchical relationships, etc., to function optimally. Lack of context can lead to misinterpretations.   | A modern data integration platform ingests and manages metadata to provide richer context and meaning to data for AI models.  |
| <b>Observability, Monitoring, and Explainability</b> | Many AI models, such as deep neural networks, are considered “black boxes” because their decision-making processes are difficult to interpret. Lack of interpretability can cause trust issues and ethical questions, especially in highly regulated industries or when making critical decisions. Lack of transparency poses challenges for observing and monitoring the behavior of AI models, which can lead to performance degradation. | Data integration tools can ensure that input data used for AI models is reliable, accurate, and representative of real-world scenarios. These tools also help explainability by providing complete visibility into where AI model data came from and what changes happened before entering the model. |

|   | AI Scaling Challenge  | How Modern Data Integration Helps   |
|---|---|---|
| <b>Integration with Existing Infrastructure</b> | AI often needs to be integrated with existing systems to be effective. This can be complex and time-consuming, particularly for large enterprises with legacy systems.  | Data integration platforms provide tools to easily integrate diverse data, allowing AI systems to securely access and analyze the needed data while respecting existing IT policies and systems.  |
| <b>Scalable Infrastructure</b>                  | Scaling AI models necessitates substantial compute resources, especially during the training and inference phases. The complexity and workload of AI models can vary, requiring dynamic allocation and optimization of resources. The challenge lies in optimizing the allocation based on the varying needs of different AI models and managing the operational costs associated with it.                                      | Modern data integration platforms facilitate the uniform distribution of data across compute clusters and cloud infrastructure. This ensures that AI models have the necessary resources for training and inference. By optimizing data storage, processing, and transfer, data integration solutions let organizations allocate resources more efficiently, manage costs, and improve the overall efficiency of AI development.  |
| <b>Governance and Regulation</b>                | The adoption of AI often raises legal and regulatory concerns, particularly regarding privacy, security, and data protection. Businesses must navigate a complex landscape of regulations such as the General Data Protection Regulation (GDPR) and ensure compliance to avoid legal consequences and reputational damage.  | Modern data integration tools are governance-ready. They provide topologies that show organizations how systems are connected and data flows across the enterprise. A centralized “mission control” console delivers deep visibility into pipelines, enabling organizations to consistently apply governance and security controls to create, process, and distribute data according to policy. They should also integrate with data lineage, governance, access, and policy control systems. |
| <b>Cost and ROI</b>                             | Scaling AI involves substantial data storage, processing, and transfer costs. As the volume of data grows, organizations face the challenge of managing these escalating costs while ensuring the efficiency and effectiveness of AI models. The costs are not just associated with hardware or cloud services but also with the operational management of data, such as ensuring data availability, reliability, and security. | Modern data integration solutions optimize data storage, facilitate efficient data processing, and minimize data transfer costs. This allows organizations to focus on innovation and development rather than operational management. It can also minimize data acquisition, storage, processing, and maintenance costs.  |

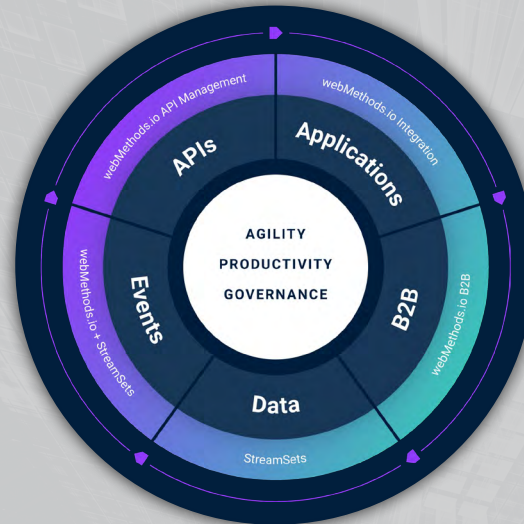
## Beyond Modern Data Integration

The right modern data integration solution provides a solid foundation for scaling your AI initiatives. It supplies the consistent, quality, explainable data AI/ML models need for reliable and trustworthy results. Other essential components include data governance and access control solutions, which the right data integration solution will support.

But you can take your foundation to the next level with an enterprise integration platform, which adds application, API, B2B, and event integration to data integration. We call this the super iPaaS and it ensures that all the data in an organization is clean, correct, and accessible for AI/ML models. It establishes a common data structure so AI systems can use diverse data types and sources. **This super iPaaS will also improve visibility into how data flows into various AI models and should have:**

- Develop anywhere, deploy anywhere capabilities so teams can work how they like and eliminate duplicate efforts
- Central control with distributed execution for faster time-to-market, simpler compliance, and better control of your integration landscape
- Closed loop app and data integration so organizations can capitalize on past, present, and future data with connectivity from apps to analytics
- A unified experience across all iPaaS components to simplify learning, managing, and collaboration across APIs, apps, data, B2B, and events
- Composable business architecture with APIs and events that gives your team a flexible set of building blocks to deliver faster
- Generative AI throughout the integration lifecycle to make the most common integration activities 10x faster, from creation to operation

It's time for a new way to think about integration  
Say hello to the Super iPaaS



A Super iPaaS finally brings together application, data, APIs, B2B, and events integrations in the same unified platform.

It is powerful enough for integration specialists, but easy enough for citizen integrators.

It is built for the future of business.

# Data Integration + AI = Enterprise-wide Success

As artificial intelligence and machine learning become more pervasive across industries, organizations must build a solid foundation to support enterprise-wide initiatives. Ensuring that AI leads to results you can trust requires ensuring the integrity and consistency of data coming into your AI infrastructure.

The right modern data integration solution provides critical functionality to overcome these hurdles and enable AI success at scale. With a focus on agility, automation, and observability, data integration streamlines and optimizes data flows to deliver high-quality, trustworthy data to AI models. With the right data foundation, AI models can deliver continuous value across the business through accurate predictions, automated decision-making, and data-driven optimization.

## Getting Started

If you're ready to build your foundation for scalable AI, the StreamSets platform provides an easy on-ramp. Data-driven organizations like Humana, IBM, GSK, and many more use the StreamSets data integration and transformation platform to rapidly deliver high-quality data for analytics, reporting, and data science.

Learn more at [www.streamsets.com](http://www.streamsets.com).



## About StreamSets

StreamSets, a Software AG company, eliminates data integration friction in complex hybrid and multi-cloud environments to keep pace with need-it-now business data demands. Our platform lets data teams unlock data—without ceding control—to enable a data-driven enterprise. Resilient and repeatable pipelines deliver analytics-ready data that improve real-time decision-making and reduce the costs and risks associated with data flow across an organization. That's why the largest companies in the world trust StreamSets to power millions of data pipelines for modern analytics, smart applications, and hybrid integration.

To learn more, visit [www.streamsets.com](http://www.streamsets.com) and follow us on [LinkedIn](#).

StreamSets  
Data Integration