tdwi

research

DATA POINTS

JUNE 2025

The Data Foundation for AI

Fern Halper, Ph.D.
TDWI VP of Research

Sponsored by

Fivetran | Microsoft

INTRODUCTION: THE MOVE TOWARDS AI, GENERATIVE AI, AND AGENTIC AI

For decades, organizations have applied AI to many different use cases, including everything from fraud detection to supply chain optimization. However, generative AI offers new and powerful capabilities, producing considerable enthusiasm. Many organizations are under pressure from executives and boards to identify meaningful applications for these technologies. For instance, in a recent TDWI survey, implementing generative AI ranked higher in analytics priorities than traditional machine learning.[1]

Yet many organizations are putting the cart before the horse. While some see early successes using generative AI—for example, improving the efficiency of writing marketing emails or assisting with software development—the true value and potential for sustained growth will depend on deeper integration with company data.

For instance, many companies initially explore generative AI by deploying simple chatbots augmented with company FAQs or documentation to answer customer questions. But to fully leverage their potential, organizations will need to incorporate dynamic, real-time company data, such as customer loyalty history, purchasing behavior, and service interactions, for personalized chatbot applications.

Another popular use case involves using generative AI for natural language processing to analyze customer sentiment from unstructured text documents and classify customer issues. These use cases can be further enriched with structured data such as billing data and behavioral variables.

Beyond generative AI, agentic AI—so called because it can exercise agency—builds on the probabilistic reasoning and language generation capabilities of generative AI. It can augment them with autonomous, goal-directed behavior and the ability to interact with external tools and systems, including retrieving relevant information through integrated retrieval mechanisms. For instance, when a high-value customer initiates a chat to cancel their service, an agentic AI system might automatically retrieve loyalty data, recent support interactions, billing history, and current promotions. It might then assess the risk of churn, select a personalized retention offer (e.g., a discounted plan or device upgrade), obtain approval if needed, and present the offer in real time.

To reach this next stage of AI maturity, organizations must establish a solid, well-structured data foundation capable of delivering personalized, context-aware insights to AI applications. Important components of a modern data foundation for AI include:

- **Integration support for all data types.** Traditional machine learning models primarily relied on structured data. However, generative and agentic AI systems require access to unstructured data such as text notes, images, audio, and video. As a result, AI-ready environments must support a wide range of data types at scale. This includes capabilities for storing unstructured data, pipelines for extracting and transforming that data into AI-usable formats, and infrastructure to support real-time data ingestion and access.
- **Ensuring data trustworthiness.** For AI to generate reliable and ethical outcomes, the data used for augmenting, training, scoring, and refining models must be trustworthy. Trustworthy data is accurate, complete, timely, and consistent; these criteria may vary depending on data type. The model inputs must be well understood, bias-free, and contextually appropriate. Trust also

[1] Unpublished 2025 TDWI Data and Analytics Survey.

extends to derived features used in machine learning and generative models.

- **Data discoverability.** Users should be able to easily locate and access the data they need. This typically involves strong metadata management, indexing, semantic tagging, and intuitive search capabilities. Enhanced discoverability accelerates innovation, reduces operational inefficiencies, and strengthens both data governance and user trust.
- **Data governance.** The enterprise must ensure data is fit for the intended purpose, which often includes being accurate, consistent, secure, and used responsibly. Data governance defines policies and processes for managing data across the organization, including who can access it, how it's classified, and how it's stored and shared. Effective governance aligns data practices with internal standards and external regulations. Modern platforms often support governance through built-in tools such as data catalogs, lineage tracking, and fine-grained access controls that make it easier to enforce policies and maintain oversight.
- **AI governance.** Managing AI systems responsibly requires oversight of model development, deployment, and maintenance. AI governance addresses issues like model transparency, explainability, regulatory compliance, and ethical alignment. It helps organizations mitigate risks related to bias, fairness, accountability, and reproducibility. Many modern platforms now include tools such as model repositories, feature stores, and monitoring capabilities to support governance throughout the AI life cycle.
- **Observability.** Organizations need to continuously monitor their data pipelines, models, and outputs to detect anomalies and track performance issues. This helps maintain trust in AI systems by ensuring that inputs and results stay accurate and reliable. Tools that support this level of monitoring—often referred to as observability—also play a key role in diagnostics, troubleshooting, and compliance tracking.

This TDWI Data Points report focuses on how enterprises are modernizing their data strategies and architectures to meet the demands of AI, navigate economic uncertainty, and address the evolving need for unified, scalable data environments. It discusses how far along many organizations are in their data foundation journey, the challenges organizations face, how organizations view the opportunities for their modern data foundation, and the value it can provide.

DATA POINTS SURVEY METHODOLOGY

TDWI Data Points uses primary research to deliver a concise view of a specific opportunity area. In May 2025, TDWI conducted a survey of senior data and analytics decision-makers. A total of one hundred and fifty-seven respondents met the quality criteria and were included in this analysis. Respondents represented a range of industries, roles, and company sizes, with the largest percentage coming from the financial services sector.

Survey responses were analyzed and evaluated across four different dimensions that form the foundation of the analysis.

- **CURRENT STATE** measures how far along respondents are in deploying their data foundation for AI, serving as a measure of maturity. Current state is categorized as early, midway, or mature.
- **CHALLENGES** assesses both implementation and ongoing challenges. Overall challenges are rated as significant, manageable, or low.
- **VALUE** examines the value organizations expect from their data foundation. Value is categorized as impactful, perceived impact, or no impact.
- **OPPORTUNITIES** examines opportunities for value and growth. Opportunity is rated as high, medium, or low.

TOP LINE DATA POINTS

Overall analyst assessment: While organizations are making some progress in building out their data foundation for AI, there is still room for improvement in order to obtain the most value from their AI investments.

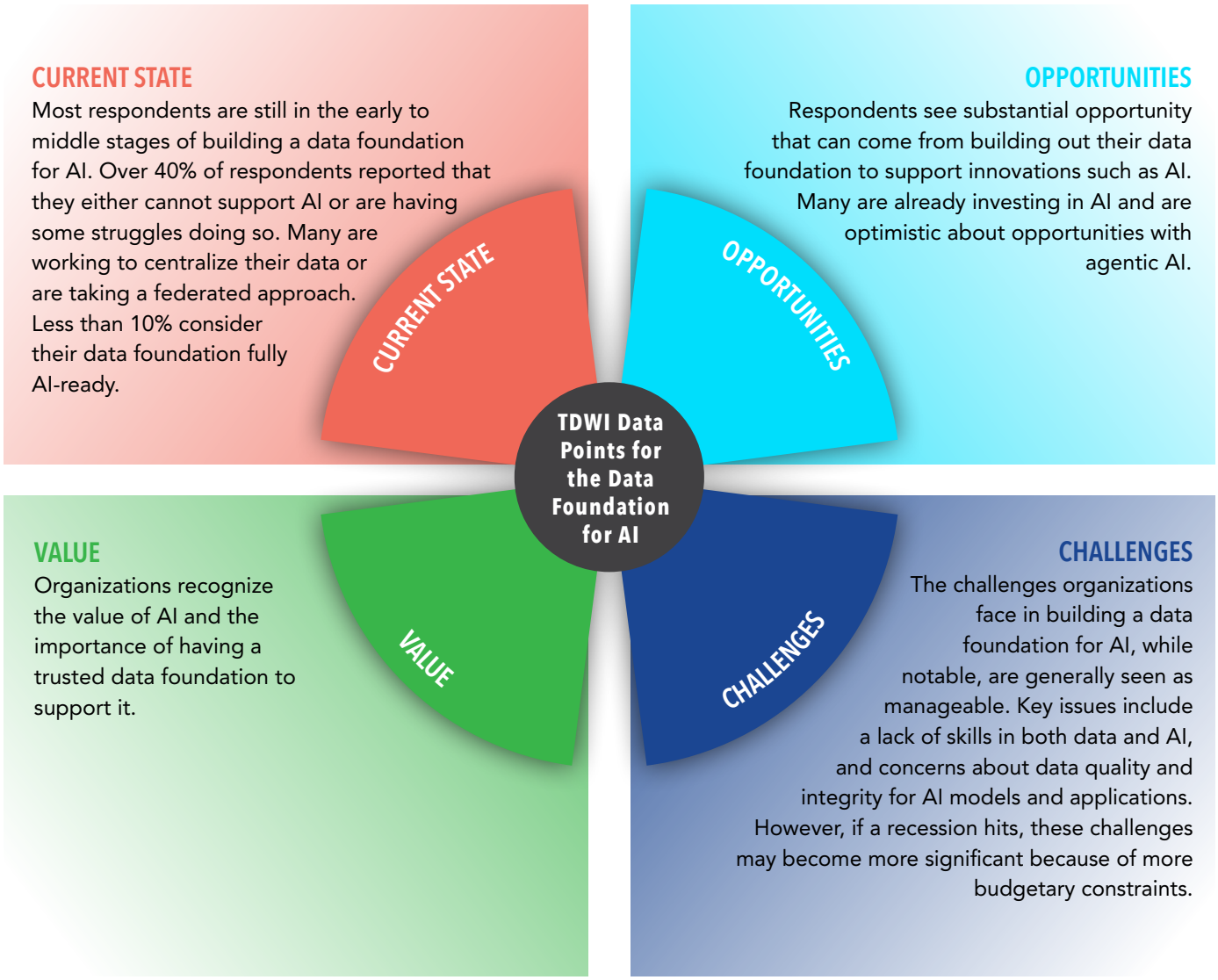

Figure 1. *Key data points for the data foundation for AI.*

RESULTS BREAKDOWN

CURRENT STATE

An AI-ready data foundation is the basis of any serious AI effort—it's a modern, scalable architecture that brings together data from across the business, makes it accessible, and ensures it's trustworthy. It supports both structured and unstructured data, enforces strong governance and security, and handles both real-time streaming and batch processing.

The Data Points survey asked respondents about the current stage of their data foundation for AI. As shown in Figure 2, 41% of respondents reported being in the very early stages of building this foundation. They are either not building a data foundation or are currently planning to build one. Twenty-three percent said they were in the process, with hybrid deployments that were only partially meeting their objectives. The rest were using either centralized cloud or federated deployments, with only 8% stating that they were completely AI-ready.

Clearly, most organizations haven't fully built the foundation that enterprise-scale AI demands. While some have made progress, few have scaled it across the business. Getting to full AI readiness requires infrastructure that can handle large volumes of diverse data and make that data easy to access and use.

These findings reflect a broader trend. TDWI research shows that many companies are still working to unify their data environments, whether through physical or logical architectures. Most are also in the middle of the journey toward building the governance needed to support responsible, enterprise-wide AI.

Figure 2. *Based on 157 completed responses.*

We also asked respondents about their approach to data storage architecture. For this survey, we defined a data lake as an object or file store that can handle large volumes of both raw, unstructured data—such as free-form text, images, or video—and structured, relational data like tables organized into schemas. Examples include Microsoft OneLake, Amazon S3, and Google Cloud Storage. For AI to be successful, organizations need to manage both types of data effectively.

The analysis reveals distinct patterns in how data storage strategies align with stages of data foundation maturity for AI. More mature organizations—those categorized as having a "Centralized AI Foundation" or being "Completely AI-Ready"—were significantly more likely to adopt modern architectures such as data lakehouses (33%–38%) compared to those in earlier stages such as "Hybrid Limitations" or "Not Building" (8% and 0% respectively, not shown). Those making use of a federated strategy are using a combination of platforms, as would be expected, and are relatively mature.

Those making use of the centralized AI foundation are often looking for a full data stack that includes AI and ML capabilities along with data integration and data governance capabilities. In this survey, 36% of respondents noted that the biggest technical driver behind their storage architecture decision was to support AI (not shown). Thirty-four percent said that it was to simplify data management and governance (not shown). We have seen this in other TDWI surveys; the top priority for modernizing a company's data foundation is to support AI.

Respondents were also asked about the current state of their AI maturity (Figure 3). Forty-six percent said they have adopted traditional AI across the enterprise, and within that group, 15% are also using generative AI with their company data. Another 39% reported using traditional AI in isolated use cases and starting to experiment with generative AI. Note that typically TDWI sees about half of survey respondents making use of AI; it is likely that the 39% who mentioned predictive AI in specific functions are early in this journey and are using generative AI in a consumerized model (see below).

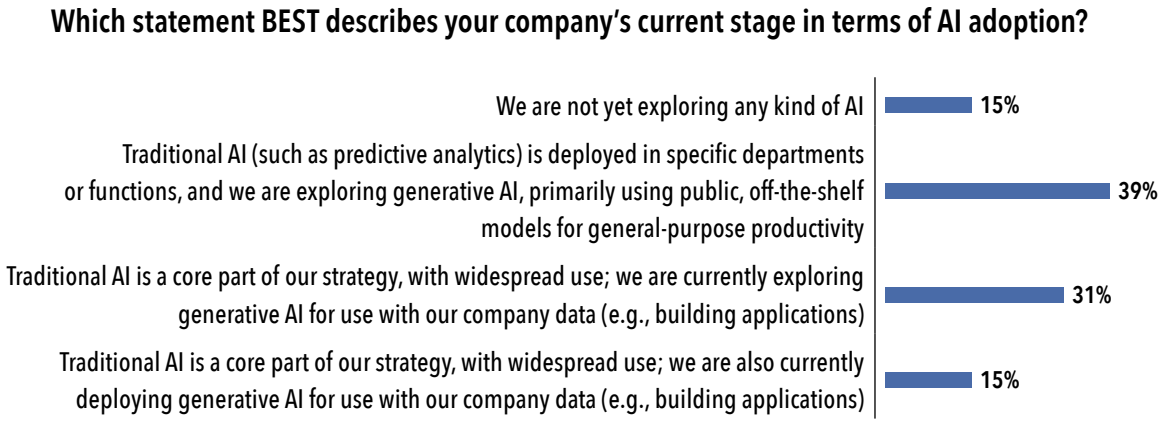

Figure 3. *Based on 157 respondents.*

A deeper analysis shows that data foundation maturity and AI adoption tend to move in lockstep. Respondents who said they were in the early stages of building their data foundation were also more likely to report limited AI use, often confined to pilots or experiments. This reinforces the idea that a modern data foundation is critical for enabling AI, not just from a technical perspective, but for overall organizational readiness as well.

However, while most organizations with advanced AI programs also report mature data foundations, there is a subset pushing forward with AI despite fragmented or partially modernized data environments. This points to a potential risk: some enterprises may be building AI capabilities on shaky ground. It highlights the need for closer strategic alignment between AI and data infrastructure teams.

This disconnect is likely driven by the rise of generative AI and early use cases that don't rely on company data. These efforts often fall into the category of "shadow AI"—tools or experiments happening outside formal IT oversight.

Taken together, these results suggest that the data foundation for AI is midway in maturity, and of course the bar keeps moving. Organizations recognize the value of robust, scalable data architectures to support AI, but many are still working through the implementation and integration challenges.

CHALLENGES

As discussed earlier, many organizations continue to face both technical and organizational challenges when building the kind of data infrastructure that AI requires. At the top of the list are foundational issues around data consistency and expertise. As shown in Figure 4, about one-third of respondents cited a lack of standardized data formats and a shortage of skills or expertise as top challenges. This points to a fundamental problem: even with the right technologies in place, many teams lack the consistent data structures and baseline knowledge required to fully leverage them.

This situation is complicated by concerns around data quality. Many respondents (31%) reported that inconsistent or poor data quality continues to undermine their AI efforts. At TDWI, we've seen that unstructured data is still widely perceived as less trustworthy than structured data. Yet without reliable data, AI initiatives are unlikely to succeed. The good news is that more organizations are beginning to adopt automated and AI-infused tools that can identify and fix issues with data quality (primarily for structured data). More organizations are also starting training programs to help teams adapt to new technologies and methodologies.

Another challenge is integration across hybrid or multi-cloud environments (cited by 27% of respondents). As mentioned earlier, many organizations operate in a mix of on-premises and cloud environments, and stitching these systems together is complex, even when an organization decides to implement a federated or data fabric approach. In fact, 23% of respondents are still relying on complex or manual transformation processes, which can slow innovation and increase operational risk. Many respondents are challenged by these issues, even those who did not report a mostly hybrid data environment overall.

In other words, while preparing for AI is a strategic priority for many companies, progress is often slowed down by a combination of legacy data issues, employee skill gaps, and architectural complexity. These challenges require the right mix of organizational and technical support for tools and training to overcome.

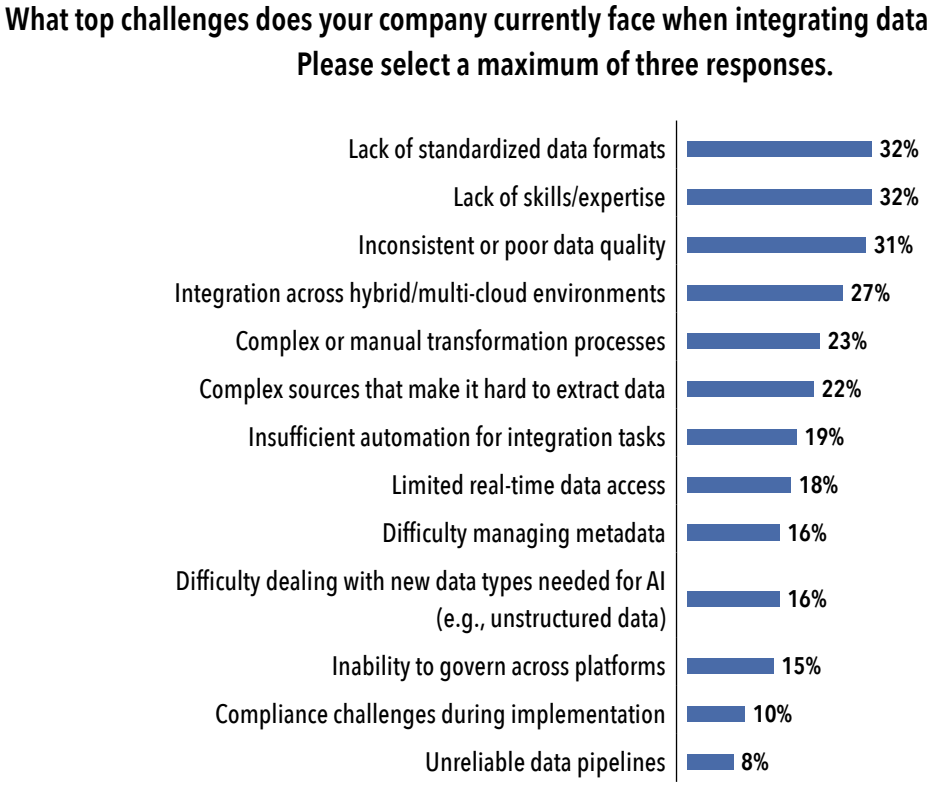

Figure 4. *Based on 157 respondents.*

While respondents ranked unreliable pipelines relatively low on the list of challenges in terms of integrating data across systems (Figure 4), they still report spending a lot of time and effort building and maintaining ETL/ELT pipelines (Figure 5). When asked about how much of their company's engineering time goes to building and maintaining data pipelines, specifically ETL/ELT, only about one-third of respondents said it was just 0–25% of their data engineer's time. The rest reported higher time investments, with the largest group (40%) indicating that ETL/ELT consumed between 26–50% of their engineer's time. This is significant. Pipeline development remains a major operational burden in practice. It continues to divert highly skilled engineering talent away from more strategic, value-driving initiatives. Automating and standardizing pipeline workflows presents a clear opportunity: reduce technical overhead, minimize maintenance cycles, and free up engineering teams to focus on innovation rather than infrastructure.

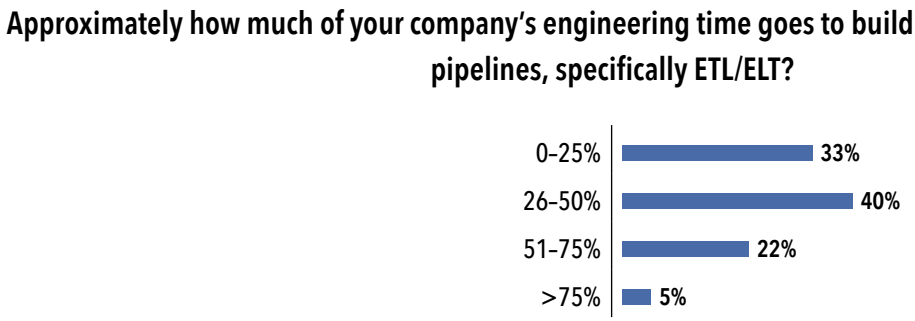

Figure 5. *Based on 157 responses.*

In addition to technical and organizational challenges, many organizations are also facing budgetary challenges. When asked about the biggest factor blocking continuing data investments, the top response was budgetary constraints (30%), followed by uncertainty around ROI (20%) and leadership priorities (18%, not shown). With an unstable economic climate, these challenges may be exacerbated.

That said, organizations remain largely committed to their data strategies. When asked how their companies have adjusted their investment in cloud data infrastructure and analytics, the top answer was to maintain investment as is (45%). Less than 10% reported they were pausing investments (not shown). This suggests a relatively strong commitment to data and AI efforts. However, if economic conditions worsen, these manageable concerns could evolve into significant roadblocks, making efficiency, automation, and demonstrable ROI even more important moving forward.

VALUE

We asked survey respondents about their planned use cases for both traditional and generative AI, as well as their current investment priorities.

For traditional AI, respondents cited a wide range of use cases already in production, including customer behavior analytics, medical diagnosis, fraud detection, predictive maintenance, threat detection, churn analysis, supply chain optimization, and forecasting. With generative AI, organizations are focusing on personalized content generation in marketing, chatbot development, and templates for customer communications. Respondents also highlighted broader goals such as enhancing overall productivity and improving operational and training processes through generative AI.

Collectively, these responses reflect a strong belief in the value potential of AI initiatives. This confidence is reinforced by prior TDWI research, which has shown that organizations that deploy AI are more likely to report measurable top- and bottom-line impact compared to those that do not.

On the data front, we asked respondents to identify which areas of their enterprise data strategy they believe present the greatest opportunity for growth or innovation for their company over the next 12-18 months. The purpose of the question was to identify where respondents felt they could derive greater value from strengthening their data foundations.

Not surprisingly, enhancing data quality and consistency to improve confidence in insights was rated significantly higher than other options (Figure 6). This emphasis on data quality directly connects to the perceived value of AI. High-quality, consistent data improves the accuracy and reliability of AI models, which in turn drives better decision-making, better outcomes, and more value for the organization.

In short, while AI use cases are expanding rapidly, the underlying data infrastructure—especially data quality—remains a critical driver of successful outcomes.

What areas of your enterprise data strategy do you believe present the greatest opportunity for growth or innovation for your company over the next 12-18 months? Please select a maximum of three responses.

Figure 6. *Based on 157 respondents.*

OPPORTUNITIES

While many organizations are still facing challenges with their data foundation for AI, most understand that a strong data architecture is the gateway to future growth. Unsurprisingly, the top area cited for growth over the next 12-18 months in this survey is generative AI (Figure 7).

Nineteen percent of respondents cited generative AI using company data as the top growth area; however, right behind this is generative AI in a consumption model (and with other kinds of AI also close in percentages). While consumption-based models offer quick access to capabilities and can accelerate experimentation, the long-term strategic value lies in generative AI trained on a company's proprietary data. This approach allows organizations to generate insights, content, and applications tailored to their unique operations, customers, and competitive context—something generic models can't provide. As companies continue to increase the maturity of their data foundations, TDWI expects to see a stronger push toward using internal data to fully realize the potential of generative AI.

Which area of your enterprise analytics strategy do you believe presents the greatest opportunity for growth in the next 12-18 months?

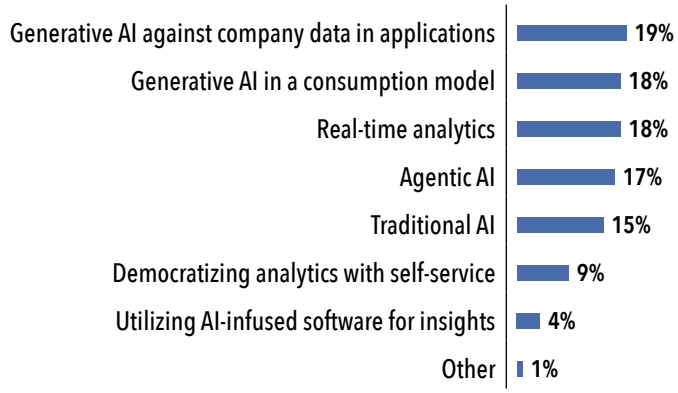

Figure 7. *Based on 157 respondents.*

While generative AI is where organizations seem to see the most growth, Figure 7 really illustrates that organizations see potential opportunities for growth in all areas of AI, certainly above areas such as democratizing analytics with self-service or even using AI-infused software to surface insights.

When it comes to investment priorities (Figure 8), AI/ML and generative AI were tied in the survey with 31% of respondents citing each as a top area of investment. Although slightly more respondents invest in generative AI than in foundational areas like data governance (where the big concern of data quality is relevant), the observed difference is not statistically significant, suggesting a growing awareness of the need to balance innovation with foundational improvements.

Additionally, while a large percentage of data engineers are spending more than 25% of their time maintaining data pipelines, only 19% of respondents selected modernizing data pipelines as one of their top two investment priorities. This disconnect suggests that budgets may be disproportionately weighted toward front-end innovation—such as AI and analytics—rather than back-end infrastructure. Or there may be a lag in recognizing the productivity return on investment in pipeline tools. Some roles who responded to the survey such as CEOs may overlook the importance of pipelines, although it puts a big burden on engineers.

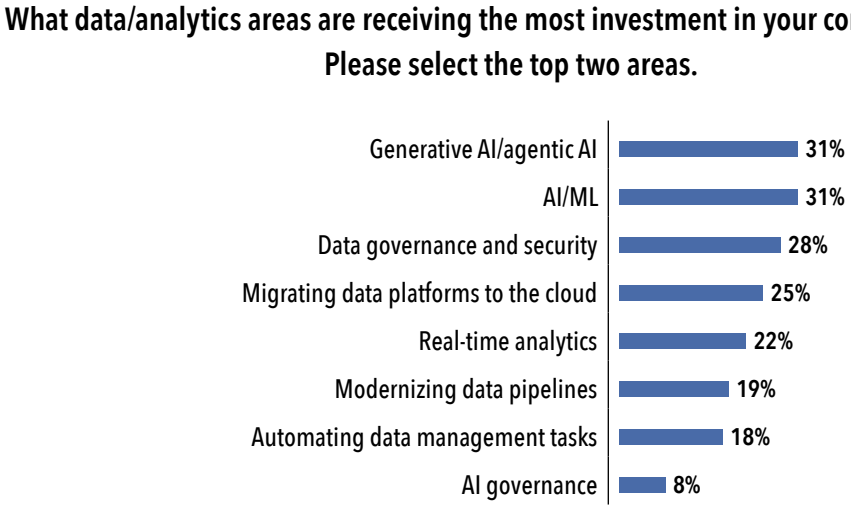

Figure 8. *Based on 157 respondents.*

In terms of agentic AI, organizations see a great deal of opportunity for task automation. When asked about the data-related tasks respondents expect AI agents to automate within their enterprise in the next 12–18 months, the top area was data quality checks and anomaly detection (57%), followed by orchestrating ETL/ELT jobs (38%), and data pipeline monitoring and alerting (37%, all not shown).

FINAL THOUGHTS: A STRATEGIC IMPERATIVE FOR DATA-DRIVEN AI

AI is reshaping the business landscape, but the race to deploy it is outpacing the readiness of most data foundations. TDWI's research shows that while AI investment is strong, true AI readiness, which includes scalable architectures, trusted data, and real-time integration, is still emerging. This disconnect threatens to undermine ROI and poses risks to organizations.

To close the gap, organizations must reimagine their data foundation as a dynamic, unified platform—not just a back-end enabler, but a strategic differentiator. The companies that will lead in AI are those that treat data as a strategic asset: continuously governed, observable, integrated, and orchestrated across the enterprise. In short, sustained success with AI will depend not only on models, but on the maturity and agility of the data infrastructure behind them.

BUILD YOUR DATA FOUNDATION FOR AI WITH FIVETRAN AND MICROSOFT

(*Content provided by Fivetran*)

TDWI's research highlights a clear message: while the pace of AI investment is accelerating, the underlying data infrastructure is not keeping up. More than 40% of organizations say they are not yet able to support AI or are struggling to do so, due largely to fragmented data systems, brittle pipelines, and limited automation. Without a solid foundation, the risks of inaccurate insights, inefficiencies, and stalled innovation only grow.

Fivetran, in partnership with Microsoft, addresses these foundational challenges head-on. Together, we help organizations modernize their data architecture with a fully managed, automated data movement platform that delivers trusted, analytics-ready data directly into the Microsoft ecosystem—Azure Synapse, Fabric, OneLake, etc. This enables enterprises to move beyond pipeline maintenance and toward driving AI at scale.

According to TDWI's findings, businesses are increasingly pursuing generative AI, agentic AI, and real-time analytics. However, they are still burdened by inconsistent data, a lack of observability, and too much time spent on manual integration tasks. This gap between ambition and readiness presents a risk—but also a clear opportunity for platform-level transformation.

Fivetran and Microsoft address these gaps by enabling organizations to:

- **Deliver high-quality, trusted data consistently:** AI outcomes depend on the quality of the input data. Fivetran's automated pipelines detect and adapt to schema drift and source changes without requiring manual intervention. This reduces the risk of broken models, data quality issues, and delays in insight delivery.
- **Unify data across fragmented environments:** As organizations balance centralized cloud deployments with federated models, they need a platform that can bridge data across SaaS applications, on-premises systems, and cloud environments. Fivetran offers more than 700 prebuilt connectors that ingest and centralize data into Microsoft destinations, ensuring accessibility for downstream analytics and AI.
- **Accelerate productivity by eliminating manual ETL:** TDWI's data shows that engineers are spending a significant portion of their time managing pipelines. Fivetran reduces this workload by up to 90% through automation and managed replication, freeing up technical teams to focus on higher-value work like building AI models or enabling business stakeholders with insights.
- **Power AI initiatives with governed, real-time data:** Whether you're deploying large language models on proprietary data, orchestrating agentic AI workflows, or enabling real-time dashboards in Microsoft Fabric, Fivetran ensures the data fueling these experiences is accurate, timely, and aligned to enterprise governance standards.

As AI investment grows, particularly in areas like generative AI, which 19% of survey respondents identified as their top opportunity, organizations must be able to reliably activate proprietary data. Fivetran's high-performance CDC replication keeps operational data in sync across Microsoft destinations so businesses can power real-time inference, generate content with GenAI, or build custom apps without lag or manual work.

In fact, a comprehensive data foundation is often the difference between isolated AI experiments and scalable, production-ready capabilities. By abstracting away pipeline maintenance and enabling data movement at scale, Fivetran and Microsoft help organizations move beyond experimentation and into operationalization.

As this TDWI report makes clear, businesses cannot afford to approach AI with siloed data or legacy integration technologies. To realize the full potential of AI, particularly with enterprise-grade generative and agentic use cases, data infrastructure must evolve in parallel.

Fivetran and Microsoft provide a proven foundation for that evolution. By automating and securing data movement across the enterprise, we help organizations reduce risk, accelerate time to value, and position their data as a competitive advantage in the age of AI.

In short, AI readiness starts with data readiness. Fivetran and Microsoft are here to make sure you're ready.

ABOUT OUR SPONSOR

Fivetran and Microsoft power enterprise-grade data pipelines that automate and simplify data movement into Azure destinations such as OneLake, Azure Synapse, Azure Databricks, and Snowflake on Azure. With 700+ prebuilt connectors, real-time CDC replication, and automated schema management, Fivetran delivers accurate, up-to-date data with 99.9% reliability and enterprise-grade security.

This fully managed solution reduces operational complexity and frees engineering teams from manual pipeline maintenance, enabling businesses to focus on innovation. Fivetran's end-to-end encryption and compliance features ensure data integrity across hybrid and multi-cloud environments.

Trusted by leading enterprises including Toll Brothers, Coca-Cola North America, Saint-Gobain, and JetBlue, Fivetran and Microsoft provide a modern data foundation that powers real-time analytics, AI, and data-driven decision-making—all on a unified platform.

Learn more about the Fivetran and Microsoft partnership here:
https://www.fivetran.com/partners/technology/microsoft-azure

ABOUT THE AUTHOR

Fern Halper, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, machine learning, AI, cognitive computing, and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. She has taught at both Colgate University and Bentley University. Her Ph.D. is from Texas A&M University.

You can reach her by email (fhalper@tdwi.org) and on LinkedIn (linkedin.com/in/fbhalper).

ABOUT TDWI RESEARCH

TDWI Research provides industry-leading research and advice for data and analytics professionals worldwide. TDWI Research focuses on modern data management, analytics, and data science approaches and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of business and technical challenges surrounding the deployment and use of data and analytics. TDWI Research offers in-depth research reports, commentary, assessments, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

ABOUT THE TDWI DATA POINTS SERIES

This series is designed to educate technical and business professionals about current trends, opportunities, and best practices focused on a specific modern topic in analytics or data management. Research for these reports is conducted via surveys of data professionals, and survey findings are supplemented by the insights and perspectives of TDWI analysts and subject matter experts. Enterprises may use these insights to overcome today's data challenges and set their strategies for the future.

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of data management and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

A Division of 1105 Media
6300 Canoga Avenue, Suite 1150
Woodland Hills, CA 91367
info@tdwi.org