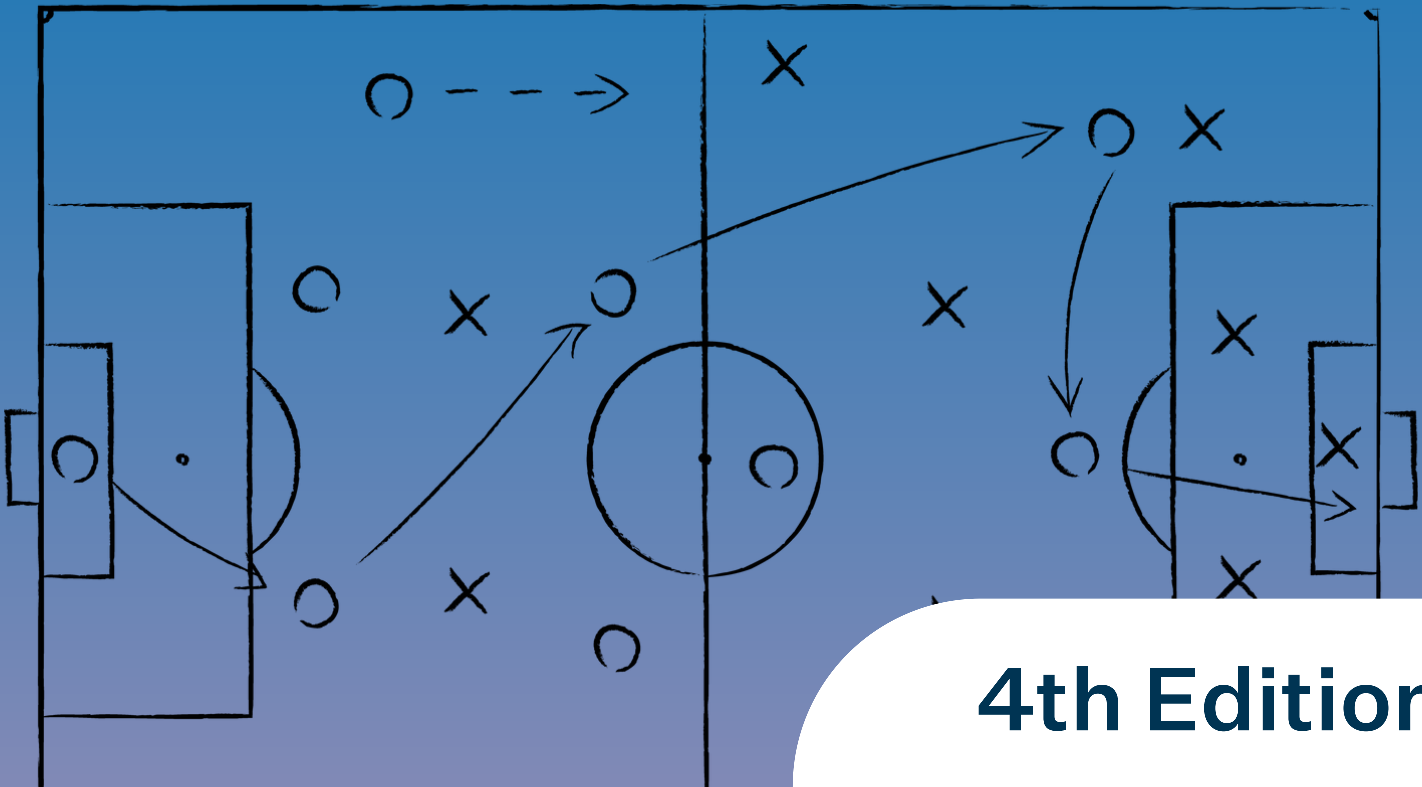
2025

MONTE CARLO

Modern Data and AI Leader's Playbook

Driving data strategy during the AI era? Learn how today's most innovative leaders are delivering stakeholder value at scale.



4th Edition

Table of Contents

Introduction

This past year, the stakes for successful data teams were raised even higher.

As generative AI continues to evolve, more organizations are finding success by putting early use cases into production. But even as AI reaches toward production, there are still significant gaps to resolve—particularly in the areas of data governance and compliance, data quality, and even realizing some actual business impact from AI.

Looking ahead, data teams will remain front and center as critical players in the AI development journey. If done right, this opportunity has the power to put data leaders in the spotlight.

If not...a searing microscope might be more likely.

So in the newest edition of our popular data leaders' playbook, we'll dive deep into the tooling essentials, growth strategies, and cultural pillars that today's teams need to succeed in the age of AI. Over the following pages, we'll share hard-won lessons—and battle-tested best practices—from data leaders who are charting the course for AI-ready data in 2025.

Let's dive in!

Section 1.
Tooling & Technologies

The Rise of the Unstructured Data Stack

According to a report by IDC, unstructured data represents about 90% of all enterprise data — but most organizations only leverage about half of that for analysis.

In the age of GenAI, that's about to change.

Enterprise success with GenAI depends on the panoply of unstructured data used to train, fine-tune, and augment models. So as more companies look to operationalize AI for enterprise use cases, enthusiasm for unstructured data will grow — and so will the need for an "unstructured data stack" designed to extract, ingest, process, and manage all this unstructured data.

For example, the data team at AssuranceIQ is exploring how they can use additional LLMs to add structure to unstructured data to scale its usefulness. They're finding early success by turning volumes of customer conversation transcripts into satisfaction scores, churn rates, and predictive models.

To get started in 2025, identify what unstructured first-party data exists within your organization — and how you could potentially activate that data for your stakeholders.

This represents a greenfield opportunity for data leaders looking to demonstrate the business value of their data platform (and hopefully secure some additional budget for priority initiatives along the way).

Enterprise AI Drives ROI — Not Revenue

Like any data product, GenAI's value comes in one of two forms: reducing costs or generating revenue.

On the revenue side, there's potential for AI-powered sales and recommendation systems to generate pipeline — but so far, these revenue-focused applications tend to produce quantity without quality.

Instead, the clearest GenAI wins have come from cutting costs by improving operational efficiency. For example, payments provider Klarna was able to drastically reduce hiring after introducing AI chatbots, while Siemens reduced unexpected failures and costs after implementing AI-powered predictive maintenance in industrial machines.

Identify AI-powered cost-cutting opportunities

Data leaders hoping to meaningfully reduce costs with AI should look for use cases that meet one of three criteria:

- Repetitive jobs
- Challenging labor market
- Urgent hiring needs

Mastering RAG Pipelines

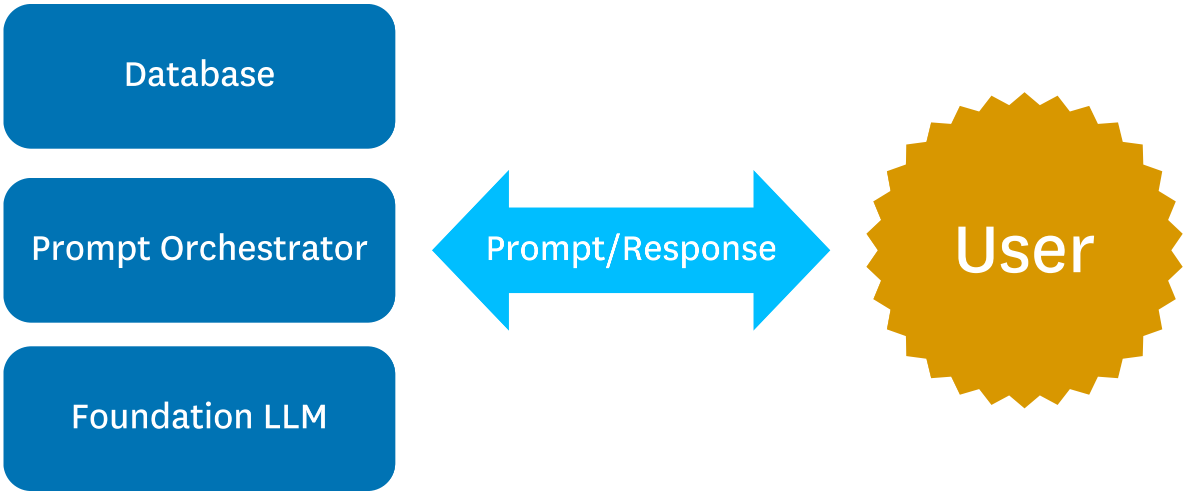
Connecting an LLM to your proprietary data is essential for nearly all AI-powered enterprise data products. In 2025, the most popular framework for getting this done will continue to be Retrieval Augmented Generation, or RAG.

RAG connects your LLM to a curated, dynamic database. This improves the LLM's outputs by allowing it to access and incorporate up-to-date and reliable information into its responses and reasoning.

Here's a simplified breakdown of how RAG works:

- Query processing: User submits a query, kicking off the RAG chain's retrieval mechanism.
- Data retrieval: RAG system searches databases to find relevant context
- LLM integration: Retrieved context is integrated with the query to create an augmented prompt for the LLM
- Response generation: LLM generates an accurate, contextually-informed and relevant response

Developing a RAG architecture is complex. But when it's done right, RAG can add an incredible amount of value to your AI initiatives.



But, RAG isn't the only game in town. Some organizations use fine-tuning, which trains an LLM on targeted datasets to improve domain-specific responses.

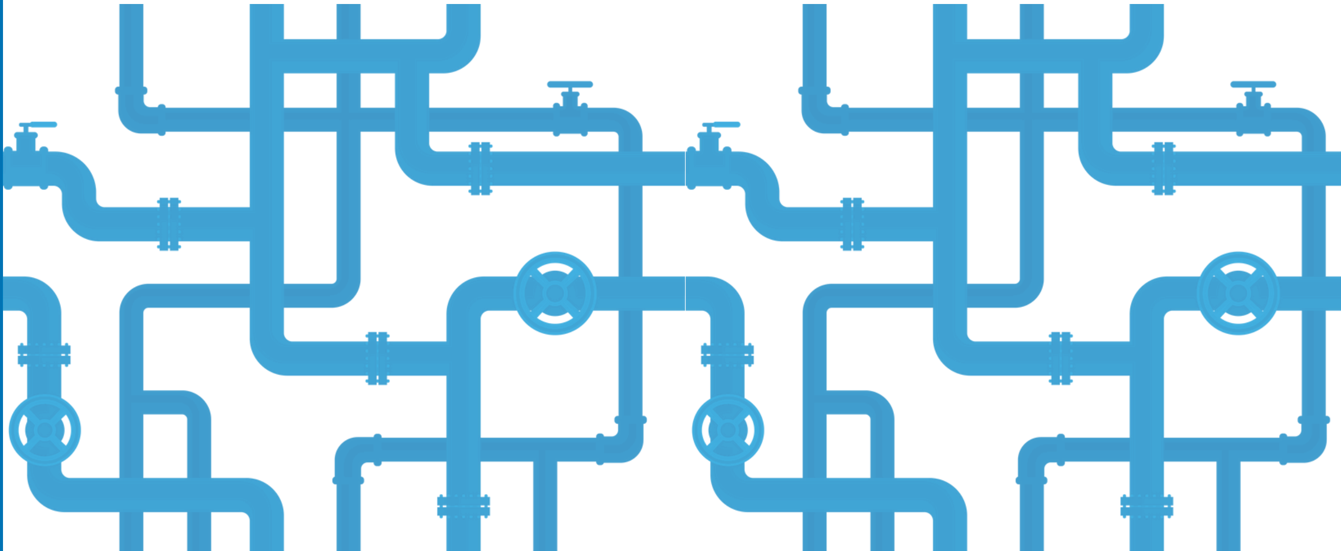
Both approaches have the same goal, but different pros and cons:

- RAG: requires more data but less compute; allows for enhanced security and data privacy; is cost-efficient and scalable; delivers trustworthy results
- Fine-tuning: requires less data but more compute; requires specialized, labeled datasets; operates as a black-box

Fine-tuning can be effective in domain-specific situations, such as responding to detailed prompts in a niche tone or style (e.g. legal briefs). For most organizations, though, RAG will be more practical, affordable, and scalable.

But here's the catch: both architectures depend on data quality at every step. If your pipelines break or schema changes unexpectedly, your entire RAG system can produce hallucinations or irrelevant responses.

Data observability is an absolute must to ensure data quality remains intact. Get both RAG architecture and data reliability right, and you'll deliver AI-powered insights your organization and customers can actually trust.



Emerging Trend: The Shift to Small Models

While the largest B2C companies will continue to use off-the-shelf models, we predict B2B environments will increasingly adopt smaller, specialized models — especially open-source. What's driving this shift?

First, small models are cheaper to run. They require significantly fewer computational resources.

Second, they improve performance. Like Google, large models are designed to answer user questions about any topic (water polo, Chinese history, French toast), so that model needs to be trained on a large enough corpus of data to deliver a relevant response. But the more topics a model is trained on, the more likely it is to conflate multiple concepts — and produce erroneous outputs over time.

Finally, ChatGPT and other managed solutions are being challenged in courts over claims their creators didn't have legal rights to the models' training data. In many cases, that's probably not wrong. Risk-averse teams may be drawn to the reduced legal exposure of smaller models, where training data provenance is clearer and more controllable.

Time will tell how this trend evolves. Large models are already aggressively cutting prices to drive demand, with ChatGPT prices already down by roughly 50% and expected to drop another 50% in the next six months. If that trend continues, big B2C players may experience a much-needed boost in their ability to compete in the AI arms race.

Emerging Trend: The Synthetic Data “Solution”

There are roughly 21-25 trillion tokens (words) on the internet right now. The AI models in production today have used all of them — yet AI requires ever-increasing amounts of data to improve. This has led to an unusual “solution”: AI generating its own training data.

As training data becomes more scarce, companies like OpenAI believe that synthetic data will be an important part of how they train their models in the future. Over the last 24 months, an entire industry has evolved to service that very vision, with companies like Tonic and Gretel generating synthetic data for specific structures or regulated industries.

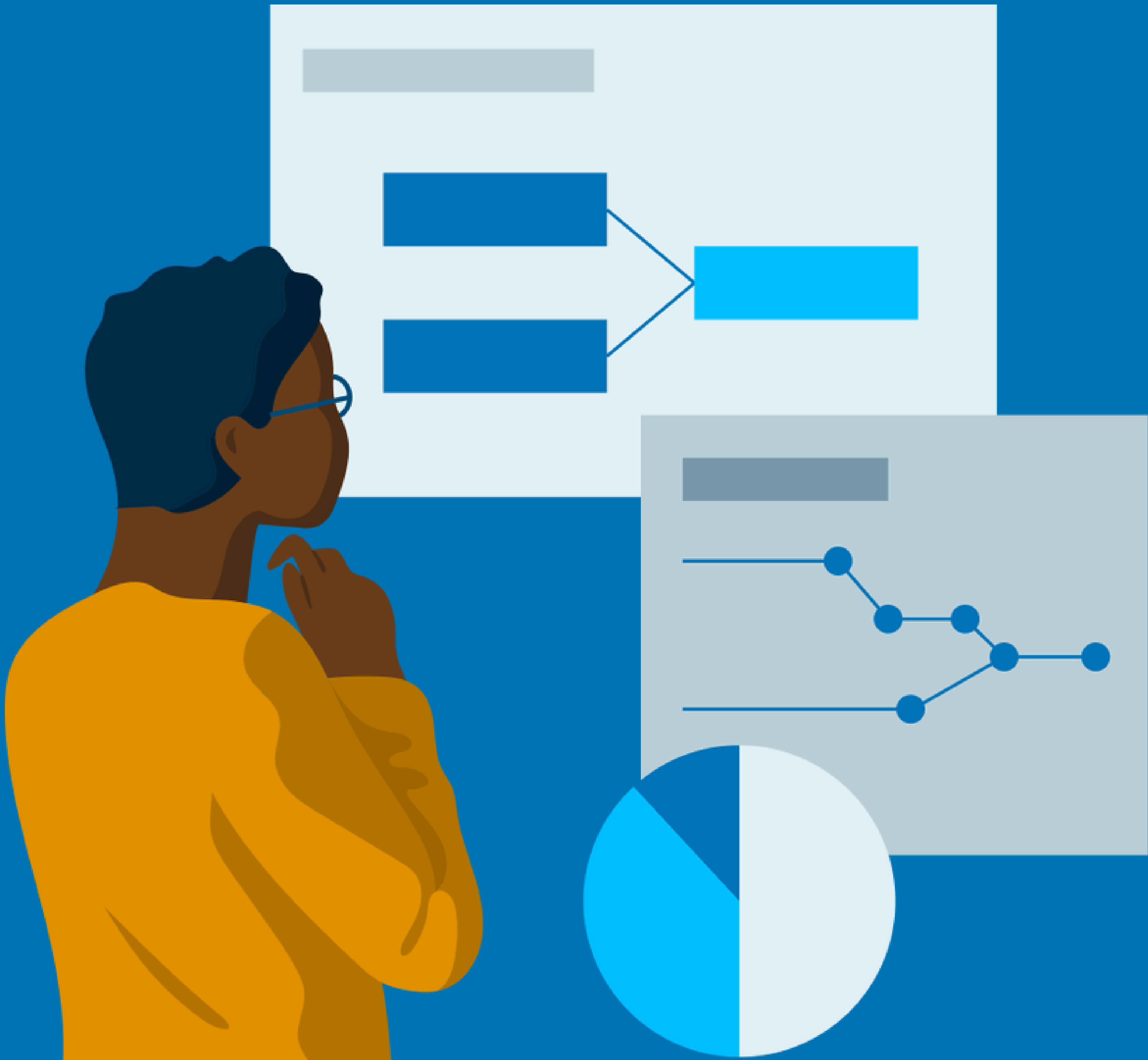
But is synthetic data a long-term solution? Probably not.

Think of it like contextual malnutrition. Just like food, if a fresh organic data source is the most nutritious data for model training, then data that’s been distilled from existing datasets must be, by its nature, less nutrient-rich. A little artificial flavoring is fine — but if that diet of synthetic training data continues into perpetuity without new grass-fed data being introduced, that model will eventually fail.

This "snake eating its tail" scenario raises important questions about AI's long-term evolution. As AI research continues to push models to their functional limits, it’s not difficult to see a world where AI reaches its functional plateau — maybe sooner than later.

Section 2.

Organizational Growth



Process > Tooling

Any new tool is only as effective as the process supporting it.

As the modern data stack has evolved over the last few years, data teams have often found themselves in a state of perpetual tire-kicking—focused too heavily on the “what” of their platform’s development, with too little focus on the “how” of implementation and adoption.

But with AI raising the stakes, figuring out how to operationalize all this new tooling is more urgent than ever.

Consider data quality. Facing the real possibility of production-ready AI, enterprise data leaders don’t have time to piece together a few dbt tests here, a couple point solutions there. They’re on the hook to deliver value now, and they need trusted solutions that they can onboard and deploy effectively today.

If you can’t get your organization up and running quickly, even the most sophisticated tools won’t amount to more than a line item on your budget and a new tab on your desktop.

Case in point: one Fortune 500 CPG leader had a best-in-class data stack to support their colossal organization, including GCP, Azure services, and columnar databases. Data played a critical role at every level of the business — inventory management, supply chain, product development, and customer engagement.

But without a scalable process in place for data quality, the data team was playing defense. Their limited resources were spent reacting to data incidents, and their time-to-resolution was Increasing. Eventually, a business decision was made based on a report with bad data — and operations suffered. It was the straw that broke the camel's back, and the data team had to transform their approach to incident management.

That meant proactive changes to process and technology, including:

- Redefining ownership: Moving data quality responsibility from engineers to analysts who work closely with the data
- Implementing analyst-friendly data observability: Selecting a data quality tool (Monte Carlo) that balanced technical capability with ease of use so the new data owners could use it

The company now uses Monte Carlo to monitor and improve data quality across the organization, with analysts quickly adopting both custom SQL and out-of-the-box asset monitoring. They now see improved incident response times, reduced engineering bottlenecks, and increased data trust across the organization.

Our prediction: In 2025, more data teams will similarly lean into proven end-to-end solutions over patchwork toolkits.

Organizations will prioritize solutions that come with established best practices and clear implementation paths so they can focus on more critical challenges like data quality ownership, incident management, and long-term domain enablement.

The Big If: AI Agents

Will AI agents transform the enterprise? Maybe — if they get off the ground. But we're not there yet.

Some background: a copilot is an AI used to complete a single step. An agent is a multi-step AI workflow that can gather information and use it to perform a task. Organizations like Github and Snowflake have seen a lot of success with copilots in 2024, but agentic AI hasn't accomplished much beyond wreaking havoc on customer support teams.

Early AI agents are an important step forward, but the accuracy of these workflows is still poor. Consider the math: state-of-the-art AI has 75%-90% accuracy, about equivalent to a high school student. But across a three-step automated workflow, that drops to around 50% accuracy overall. This makes most current AI agents too unreliable for production environments, despite impressive demos and venture capital enthusiasm.

Until the accuracy problem is solved, agentic AI's practical impact will remain limited. So despite the hype around Silicon Valley, data leaders should remain skeptical.

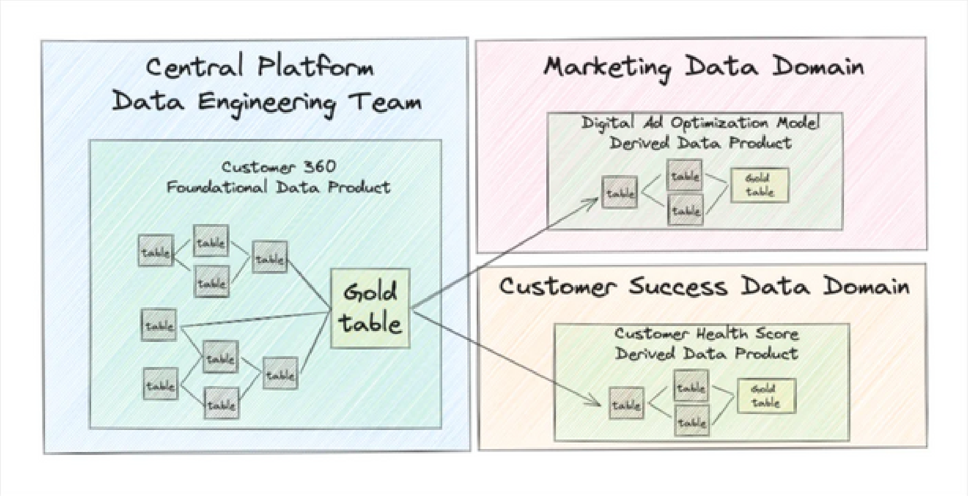
The Operationalization of Data Reliability

Managing data quality in the enterprise is a marathon and a relay. And just like any relay race, every time the baton is passed, there's a new opportunity to drop it. So defining data ownership — even across multiple domain layers and hundreds of interconnected data products — is essential to implementing good data quality processes.

Here's a framework to help you chart the course:

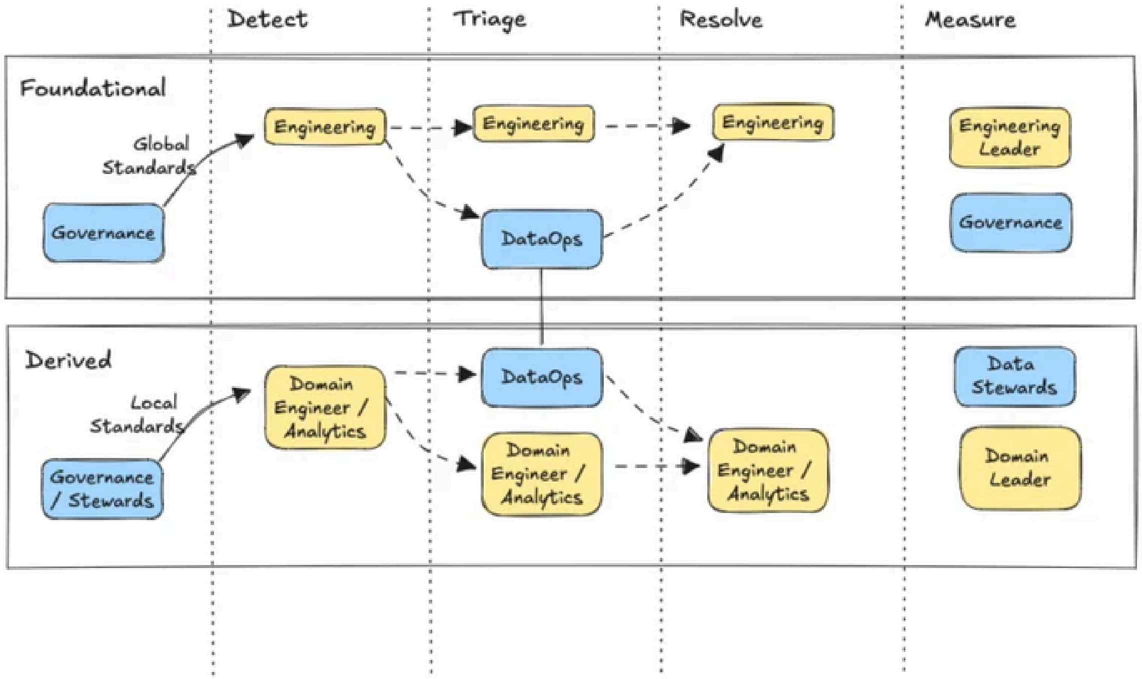
1. Identify your most valuable data products. You can't (and shouldn't) monitor every table for every issue all the time.
2. Differentiate between foundational and derived data products. Foundational products are owned by a centralized platform team and designed to serve hundreds of use cases across the business. Derived products are built on top of foundational data products for a specific use case, and owned by that domain-aligned data team.
3. Divide data quality management responsibilities by domain and product type.

For example, a "Single View of Customer" is a foundational data product that feeds derived products like an upsell model, churn forecasting, and an enterprise dashboard.



With clear ownership, you can enact different processes for detecting, triaging, and measuring data quality incidents — dependent on the foundational vs. derived.

- Detection: Domain-aligned data stewards and analysts handle monitoring and baseline data quality
- Triage: Dedicated domain DataOps or triage teams to support products within that domain
- Resolution: Domain-aligned data engineering uses data observability that connects anomalies to root cause
- Measurement: Data stewards managing domain-specific SLAs



In 2025, this structured approach to data quality ownership, combined with modern tooling and clear processes, will help organizations finally move from reactive to proactive data reliability management.

Section 3.
Culture

Blurring the Analyst/Engineer Lines

Scaling pipeline production usually leads data teams to run into two challenges: analysts who don't have enough technical experience and data engineers who don't have enough time.

Sounds like a problem for AI.

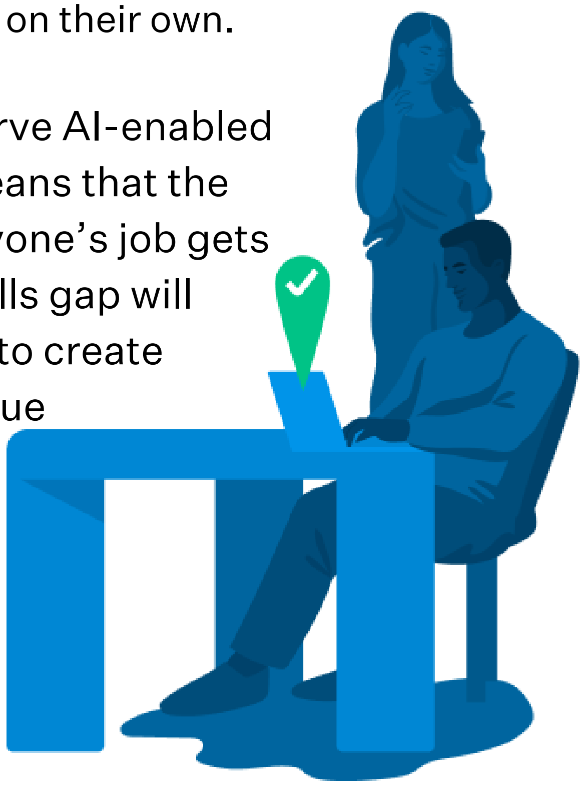
We foresee two major developments driving the consolidation of engineering and analytical responsibilities in 2025:

- Increased demand: As business leaders' appetite for data and AI products grows, data teams will be on the hook to do more with less. In an effort to minimize bottlenecks, leaders will naturally empower previously specialized teams to absorb more responsibility for their pipelines — and their stakeholders.

- Improvements in automation: New demand always drives new innovation — in this case, AI-enabled pipelines. As technologies naturally become more automated, engineers will be empowered to do more with less, while analysts will be empowered to do more on their own.

The move toward self-serve AI-enabled pipeline management means that the most painful part of everyone's job gets automated away. The skills gap will decrease, and the ability to create and demonstrate new value expands in the process.

Cheers to that future.



Third-Party Data Gets Scarier With AI

Data leaders have a love-hate relationship with third-party data — emphasis on the hate. While third-party data can power critical products and give a competitive advantage, it introduces new risks. Because regardless of where data comes from, the moment it lands in production, it’s the data team’s responsibility.

Yet again, AI is adding urgency to solve this problem. According to a recent survey, while 91% of data leaders are building AI applications, two-thirds don't fully trust their training data, much of which comes from third-party sources.

Here are practical steps to ensure third-party data quality at scale:

- Map dependencies between third-party sources and downstream products, along with all the transformations in between. Bonus points for automating it.
- Assign owners to create accountability and speed up incident response.
- Automate rule creation to scale monitoring across third-party sources.
- Integrate incident workflows with existing communication tools, like Slack or Microsoft Teams.
- Measure performance over time to understand what third-party sources break frequently and flag coverage gaps on downstream tables.

You may not be able to prevent every third-party data issue, but you can take proactive steps to detect and resolve problems before they impact the business.

Cracking the Data Quality Scorecard

Incentives and KPIs drive good behavior — it's why nailing down a sales compensation plan often warrants a spot on the board meeting agenda. In 2025, data leaders should give the same attention to data quality scorecards.

Launching an effective data quality initiative isn't easy. But we've seen these best practices make the difference between program success and having another kickoff next January:

1. Know what data matters. Talk to the business and figure out what their most critical data assets are. Start a discussion on requirements and priorities, prove out the concept of a data quality initiative that meets their needs, and then worry about scaling.
2. Measure the machine, not just the data itself. Beyond the traditional six dimensions of data quality (validity, completeness, consistency, timeliness, uniqueness, accuracy), you'll want to evaluate system reliability, usability, and operational response. This includes monitoring coverage, freshness SLAs, schema management, and incident response times.
3. Use carrots and sticks. Your stick: a minimum set of requirements for data to be onboarded onto the platform. Your carrot: a much more stringent set of requirements to be certified at each level. Once consumers can easily spot the difference, they develop a taste for highly reliable data — and producers will rise to the occasion their data actually gets used.
4. Automate evaluation and discovery. Scoring criteria should be measured automatically and immediately understandable. The most common ways to achieve this are with data observability and quality solutions, and data catalogs.

Data Products Are Having Their Moment

Data products are back this year — bigger, better, and with a lower barrier to entry than ever.

After riding the hype cycle in 2022 and falling into the trough of disillusionment in 2023, data products have come into their own in the era of LLMs. And while early definitions focused narrowly on analytics, today's data products span a broader spectrum — from self-serve dashboards to full-fledged digital services that directly generate business value.

This evolution is thanks to companies finally realizing the value of their modern data stacks, including:

- No-code/low-code platforms that have democratized data product creation, allowing business teams to build solutions without deep technical expertise.
- Data marketplaces that have emerged as central hubs where teams can discover, access, and share data products across the organization.
- Enhanced BI customization capabilities that allow data products to be more precisely tailored to specific business needs.
- More data in the cloud creating an environment where data products can be developed, deployed, and scaled more efficiently.

If your team hasn’t been exploring the possibilities of building data products, it’s time to update your roadmap for 2025.

Conclusion and Predictions for the Future

Across this playbook, a few recurring themes emerge:

- The inherent uncertainty to these evolving technologies — there's a lot of unavoidable "let's wait and see" for data leaders as AI development plays out.
- The need for identifying and ruthlessly prioritizing how to make the biggest business impact with any AI or data initiative.
- Maintaining data quality is the biggest opportunity — and a significant challenge — for data leaders in the age of AI.

But there are a few additional predictions we can make with some degree of confidence, if not certainty:

- AI will bring businesses closer to data.
- The data lakehouse, in all its unstructured glory, will be foundational for AI.
- Data teams will be more important than ever, so get used to that spotlight.

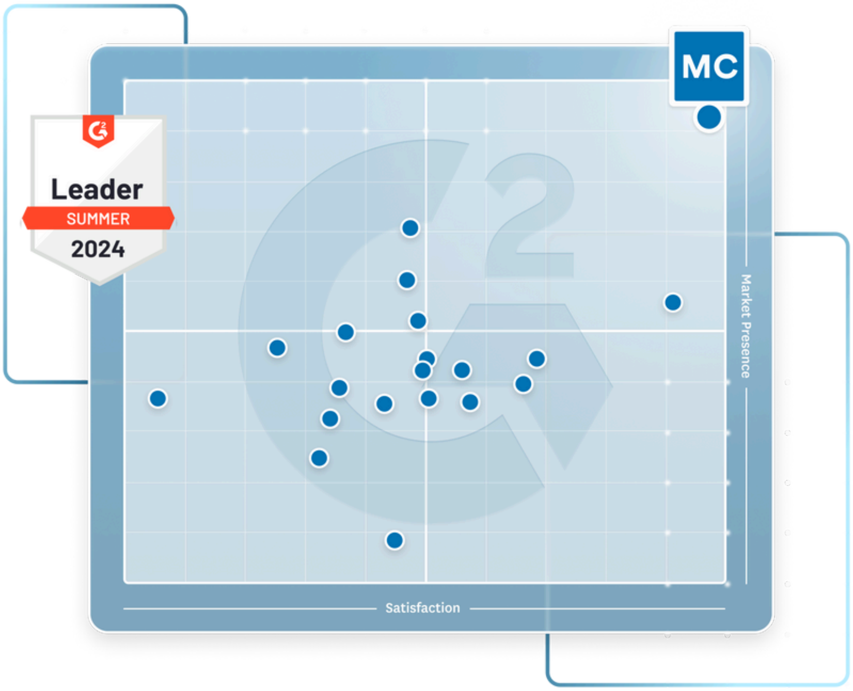
Additional Resources

Don’t let this be the ending point of your data quality journey! Check out more helpful resources including:

- Data Downtime Blog: Get fresh tips, how-tos, and expert advice on all things data.

- The Modern Data Quality and Data Observability Guide: A comprehensive guide to uncovering the best data reliability solutions for your organization.

- Request A Demo: Talk to our team to get a more accurate assessment of your data downtime, its costs, and what level of value you can expect from Monte Carlo.


Leader
SUMMER
MC
Market Presence
Satisfaction