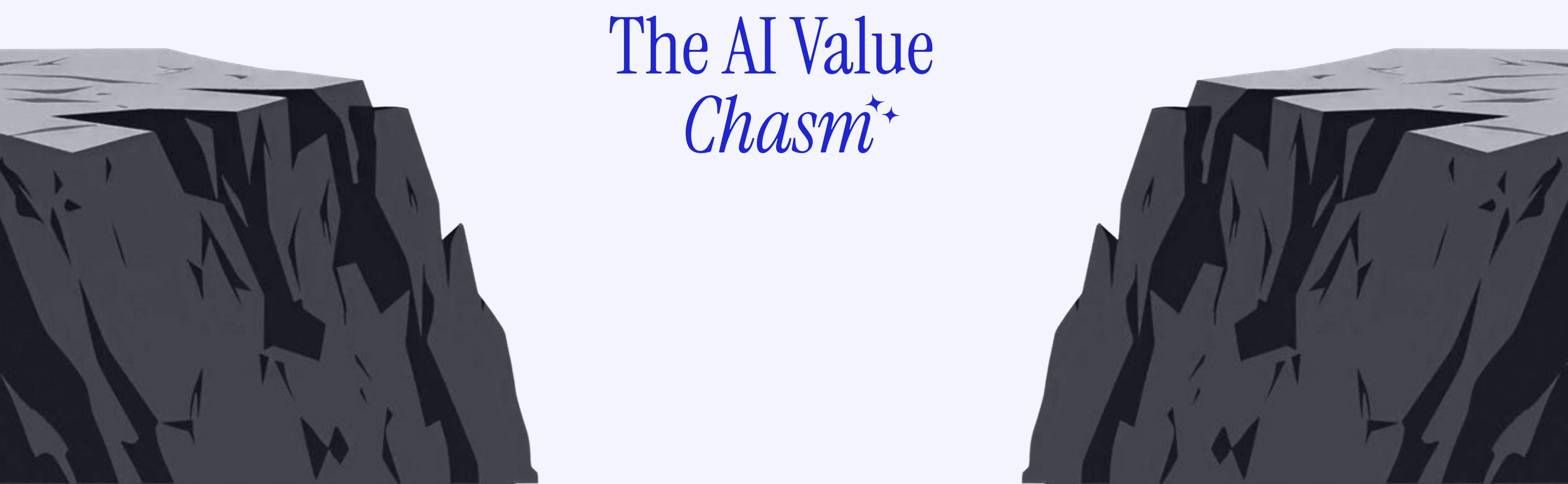
# Introduction

Every year feels like it moved faster than the last one, but 2025 was something else entirely.

AI isn't new; after all, ChatGPT has been around for years. But this year, reasoning models dropped and changed everything. In contrast to LLMs that provide answers based on pattern recognition (from a massive corpus of training data), reasoning models follow a step-by-step logical process to solve complex problems. This led to a "sudden and significant improvement" in AI capabilities, where new reasoning models like OpenAI o1, GPT-4o, and DeepSeek-R1 blew away previous benchmarks, scoring above 90% accuracy in some tests.

Overnight, AI went from something fun to experiment with on the weekend to something that seemed like it could take on real business problems. Funding surged, enterprise adoption spiked, and vendors rushed to rebuild their roadmaps around the new promise of AI.

However, just a few months later, the hype was cut short when a viral MIT report found that **95%** of AI pilots are failing. We saw the same thing in our survey of 550+ data leaders this year: **84%** of organizations said that they're betting big on AI, but only **17%** of them have actually made it a success. This is what I call the AI Value Chasm, the gap between the exciting promise and delivered value of AI that defined this year.



The models are getting better and better, but they're still failing… What does this mean for the data world?

Plenty have argued in the last couple of years that the modern data stack is failing. We've seen the data world as we know it — heaps of structured data stored in closed data warehouses, ushered through ETL pipelines by data engineers, landing in the hands of BI analysts for yet another dashboard — bending under the inevitable weight of new business and AI demands. Meanwhile, the data space is shifting fast with major acquisitions and changes. Everyone agrees that the stack has to evolve, and fast.

Here's my quick take: if the last decade was about building the modern data stack, the next one will be about rebuilding it for a world where AI is not just a feature on top of data, but a major consumer, interpreter, and activator of it.

Despite all the pessimism and doubt, I think we're moving toward a data stack that can use the power of AI in a real way for real business problems. Andreessen Horowitz named the AI-native data stack as one of their Big Ideas 2026, saying "We're excited by ways AI can continue to transform multiple parts of the data stack, and we're beginning to see how data and AI infrastructure are becoming inextricably linked." This won't be effortless though — AI deployments are still blocked by a lack of "AI readiness" in data context, governance, security, and more.

**In this resource, I'll break down my seven predictions for the modern AI-native data stack in 2026, based on research and conversations with hundreds of data leaders and practitioners.**

prukalpa

**Prukalpa Sankar**
Founder and Co-CEO
Atlan

# Analytics will be fundamentally reimagined around unstructured data

For decades, structured data has been at the center of analytics. Analysts turn the messy real world into neat tables, because ultimately we can't observe everything that goes into those datasets. Structured data and metrics let us capture and transform infinitely complex events in a few data points.

Take websites as an example. Metrics like conversion rate are valuable because there's no way for analysts to watch every person who visits a website. These stats are powerful but also limited. We can tell what number of website visitors aren't converting, but why? Is the "Book a Demo" button too hard to find? Is the product copy's tone turning people away? Are they looking for a feature that isn't available? This is when the intimidating and often infeasible task of qualitative analysis on unstructured data comes into play.

When we compress human interactions into a handful of columns, we gain valuable insights but also lose nuance. This is an inevitable tradeoff in traditional analytics, but AI changes the ground rules. Humans can't watch every website interaction, but with enough computing power, AI can.

This isn't a trivial task. In Huble's AI Data Readiness Report from April, **45%** of companies reported that using "unstructured, fragmented data" with AI was a challenge. After all, unstructured data is notoriously difficult and requires more complex data storage and processing. But its value can't be ignored: on average, business data is growing by **63%** each month, up to 90% of which is unstructured data. Andreessen Horowitz's Big Ideas 2026 called the potential to finally work with unstructured data a "generational opportunity."

But there are signs that this technology is quickly maturing. Snowflake released a paper about how their Cortex AISQL can handle declarative querying over mixed structured and unstructured data at scale. Similarly Microsoft, Google Cloud, and Salesforce have released features to work with unstructured data. Companies have already reported using generative AI to create customer support agents from an internal knowledge base, develop new product ideas from product feedback and market research, and use call center transcripts to identify common topics, improve customer service training, and forecast call volumes for different customer segments.

This evolving convergence of structured and unstructured data is one of the reasons we launched the Atlan App Framework in 2025, and saw partners and customers quickly start building an ecosystem of apps to push and pull context to Atlan from any source.


## MY PREDICTION

In the near future, I expect analytics to be reimagined from the ground up around unstructured data — not as a nice-to-have but a primary signal. Instead of just reporting what happened, we'll be able to better understand why. Metrics will never go away, but they'll likely become more complex and richer over time.

As McKinsey said, every enterprise that wants to be data-driven will soon need to "query and understand relationships between unstructured and semistructured data easier and faster". This will fuel a proliferation of new generative models powering use cases around vast unstructured data like support chat logs, meeting transcripts, compliance filings, surveillance video, medical records, and legal documents.

I think this will also fuel a broader change around how we work with data. As unstructured data becomes a first-class citizen, we'll encounter a new wave of challenges around how to reliably capture, process, and feed it into LLMs. We'll see ingestion, cleaning, and transformation shift from human-centered workflows to more automated systems where LLMs automatically capture unstructured inputs, enrich them, and pass them to humans as needed for verification. Just as today's data pipelines were built around the earlier explosion of structured data, we'll see a new generation of connectors, pipelines, and analysis tools built for the specific challenges of analytics on unstructured data.

# Data "agents" will go mainstream, freeing up time for higher impact work

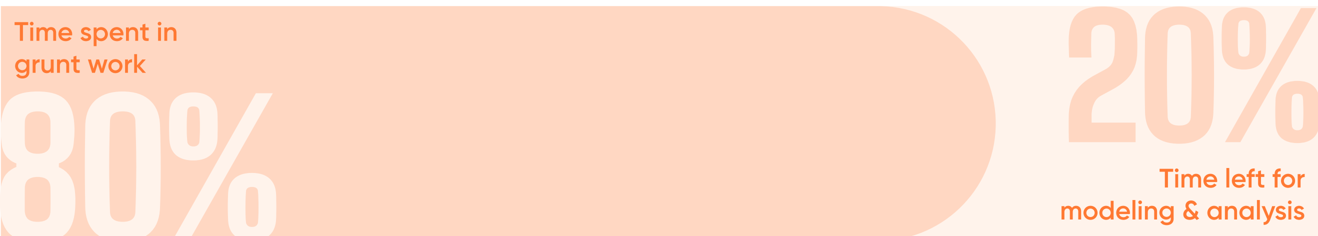
Most data teams today are still burdened by manual, repetitive tasks like writing SQL, maintaining ETL pipelines, diagnosing broken jobs, or cleaning messy data. This grunt work takes up to 80% of their time, leaving only 20% for the modeling and analysis that matters most.

It's no wonder that, according to our survey, 36% of organizations say they've built AI that works in pilots but fails to gain real traction. Without automation of the basic grunt work, most teams simply don't have the bandwidth to operationalize AI.

Thankfully, though, AI is already coming for these important but tedious tasks.
There are three main types of agentic workflows emerging today:

- **AI-assisted workflows:** Think of tools like Cursor for software engineering, which accelerate work while humans remain in full control.
- **AI-augmented workflows:** Think of tools like Devin, where the AI does most of the work and consults humans only when needed.
- **Fully autonomous workflows:** This isn't a reality yet, but we may see AI agents owning entire tasks end-to-end (without human intervention) in the future.

2025 saw the launch of generative coding tools from major LLMs (Codex, Claude Code), core data products (Databricks' Mosaic AI, Snowflake's Cortex Assist, GitHub's Copilot) and splashy startups (Cursor, Devin, Ciroos) alike. According to recent data from ChatGPT, coding and data analysis are among the top use cases for enterprise AI today.



Assisted workflows are already happening in the wild — long before most companies have formalized agentic tooling. Data people are already using AI-assisted workflows to help people find the right data, write and auto-debug SQL code from natural language queries, add new columns, build pipelines, carry out root cause analysis, and more. Early in 2025, we were the first metadata platform to launch an MCP Server, which is being used by customers to power agentic discovery and governance, and this is only the beginning.

With simpler tasks already in sight, some people are testing the waters with more complex use cases. For example, Databricks published a paper showing that LLM-generated data contracts can reduce manual effort by over 70%, and a group of data leaders outlined a Digital Data Steward built on agentic AI. Gartner even predicted that by 2028, at least "15% of day-to-day work decisions will be made autonomously through agentic AI" and "33% of enterprise software applications will include agentic AI".

## MY PREDICTION

We'll shift rapidly from today's AI-assisted workflows to more autonomous data-centered workflows, where data agents take on meaningful analytical and engineering tasks. Data people who continue with "business as usual" — writing every unique SQL query and building every one-off dashboard — will soon be overshadowed by those who learn how to hand off their manual work to AI and focus on higher value work.

**There are some key data agents I'm excited about:**

- **AI data steward:** Autogenerates documentation by reading through SQL queries, other code, and existing context from the data estate
- **AI root cause analysis agent:** Analyzes lineage when something doesn't look right in a metric or dashboard
- **AI data customer service:** Resolves business users' tickets (thus replacing the existing Jira or self-service support workflow at many organizations)
- **AI data engineer:** Automates painful data engineering tasks like adding a column to a table.

A key issue though is, what context do you need to feed these AI agents?
Just like a new data analyst, there's only so much that an AI agent can do without proper onboarding and training. This is where the context layer comes into play — more on that soon!

# The AI analyst will finally make the self-service dream a reality

For years, the data world has dreamed of self-service analytics — business people able to query metrics, generate insights, and create dashboards on their own, no data team needed. Tools like Tableau and PowerBI and Looker were supposed to solve this, but a decade later, we just ended up with embarrassingly low adoption and dashboard fatigue.

Recently, conversational data analytics via chatbots and Slack agents have been gaining traction. Existing BI companies are making a play for the space, like Tableau launching Pulse, Hex's Threads, and Thoughtspot's Spotter. We've also seen a wave of activity from other data companies (e.g. Snowflake's Intelligence, Databricks' AI/BI Genie, Dagster's Compass) and new companies (e.g. Basedash, Julius, Scoop, Wisdom AI, and Wobby). These tools are promising but early, with adoption often limited to quick experiments, ad-hoc queries, and small reports.

The gap with self-service analytics has always been **data literacy.** Business leaders know the ins and outs of their company, but they can struggle to use the right data in the right way. In contrast, conversational analytic tools speak the language of data, but they don't know each company and industry's unique lingo. I think that mature AI will bridge this gap.

## MY PREDICTION

I think that we'll finally unlock self-service BI, where anyone can "talk to the data", in the near future.

This development is a work in progress. We are currently in the midst of the hype cycle where every data team wants their business to "talk to their data" — at the Peak of Inflated Expectations, about to enter the Trough of Disillusionment. People are beginning to realize that it isn't easy to get AI to talk to their data, and going from demo to production is a lot of work. That being said, I do think that we are at the precipice of something magical.

For the longest time, data people have been trying to teach business the language of data. Now it feels like we are ready instead to teach data the language of business, thanks to large language models. With AI agents that can understand both business language and data semantics (with the help of a context layer — more on this next), I think we'll see AI stepping into more business use cases.

We've already seen this working with real customer applications in Atlan's AI Labs. We've worked with enterprise customers to build iterative loops that leverage context in Atlan to optimize the semantic view used by their AI analysts, increasing the accuracy of their responses by **5x.**

Whether this replaces traditional BI entirely is still up in the air, but I think it will make self-service analytics more possible than ever before. The question of who will own this, though, is still an open question since we've already seen data warehouses, BI tools, AI companies, and newer startups making a play for this space. However, I do think the eventual user pattern for where an AI analyst will live will likely be a single conversational interface in businesses and connect to AI analysts via MCP or other agentic protocols (e.g. Google Gemini, Microsoft Copilot, and ChatGPT Enterprise).

# The context layer will become foundational to enterprise AI

In our survey of data leaders, there was one clear message: AI doesn't know what it needs to know to be successful. Nearly half (**49%**) of AI failures were attributed to a "lack of business context", and **26%** cited a "lack of trust in AI output" and **45%** cited hallucinations as a blocker to scaling AI.

This reinforces the disconnect between AI hype and reality — the AI Value Chasm that I mentioned earlier. While **84%** of organizations said that they're betting big on AI, **59%** are stuck in planning or limited pilots, according to our survey. And many of these pilots aren't even going anywhere. The majority (**54%**) of organizations have deployed nothing or only "a few small-scale launches", while only **5%** have deployed "nearly all" of what they built.

This is the AI context gap at work. When data lacks context, AI can't understand the specifics of a company's industry, jargon, structure, and edge cases. Think of the thousands of unwritten rules at your organization that people just have to learn: when to escalate, what that specific executive means by "the latest revenue metric", or what dashboard to never use.

Even when context exists, that can create its own problems. We heard from enterprises who have implemented AI across verticals, setting up everything from hiring automations to threat detection. It's possible to train each agent or application with enough context to do its job fairly well, but those tools can't talk with one another. A sales chatbot may be able to talk about pricing or a support chatbot about product issues, but neither chatbot has the full picture. That's a problem when a user inevitably asks the wrong thing in the wrong place, and AI is left without an answer. At the same time, enterprises have realized that by uploading specific context into each product, they're locking themselves long-term into a fragmented network of individual AI tools.

Tobi Lutke and Andrej Karpathy framed this problem as "context engineering". As Andrej explained, "People associate prompts with short task descriptions you'd give an LLM in your day-to-day use. When in every industrial-strength LLM app, context engineering is the delicate art and science of filling the context window with just the right information for the next step." This problem is exacerbated by context rot, where Chroma found that across 18 LLMs, "model performance grows increasingly unreliable as input length grows".

Companies have already started focusing on the semantics problem. For example, Palantir's emphasis on semantics with Ontology has been key to its massive recent commercial momentum. We at Atlan along with Snowflake, Salesforce, dbt Labs, and others launched the Open Semantic Interchange initiative, which aims to accelerate AI and BI adoption by moving fragmented data definitions into a universal semantic data framework. The launch of SAP Business Data Cloud with Databricks included a focus on semantics, including the ability to make data searchable and actionable via a knowledge graph of business metadata and semantics. Microsoft announced Fabric IQ, a unified intelligence platform powered by semantic understanding and agentic AI.

## MY PREDICTION

I think that what the data warehouse layer is for BI, the context layer will be for AI. Everything in the near future will come down to providing the right context and governance to AI, and companies will soon realize that their "alpha" in the AI world is open, interoperable context that can interplay across a variety of agentic and AI use cases.

I think that soon we'll see context and semantics treated as infrastructure: a shared, federated foundational layer for enterprise data and AI. As Tomasz Tunguz explained, "Semantic layers told AI what data meant. Analytical context databases teach AI how to reason about it."

I personally call this the "context layer", which encodes how a company thinks, decides, and acts. It should turn scattered, static knowledge into a single, reliable frame of reference, i.e. a living map of how the organization thinks and operates. This relies on creating systems that capture, maintain, and deliver context automatically through four key capabilities:

- **Context extraction:** automated ingestion of metadata, lineage, business definitions, usage patterns
- **Context products:** "minimum viable product" curated bundles of context (e.g., "customer context," "sales context," "product context")
- **Human-in-the-loop context feedback loops:** where analysts, domain experts, and business users can enrich and correct metadata to build shared meaning
- **Context store:** a unified, versioned store of context that AI agents, BI tools, and humans alike can query

# The separate data & AI stacks will fuse together into a new AI-native data stack

In recent years, the modern data stack (think ingest > storage > transform > BI) and the "AI stack" (models > agents > application) have evolved somewhat separately. We've seen data vendors start to integrate AI into their products, but the core AI tools are seen as a parallel but separate universe.

In our survey, **34%** of data leaders reported that AI gets siloed with limited access rather than being fully integrated into their tech stack. However, I think this is about to change.

In the last year, Fivetran and dbt Labs merged, signaling a unification between deep transformation and modeling. Similarly, Salesforce acquired Informatica, bringing data integration and metadata management to the Salesforce platform. And a few months before that, Databricks championed a vision of unified data, AI, and governance in a "seamless, open ecosystem" at their Data + AI Summit 2025. We've mostly stopped arguing about warehouse vs. lakehouse and aligned on open storage and cataloging that can be used by AI and data infrastructure alike — which are becoming increasingly intertwined as AI leaks into traditional "data" tasks and technology, and vice versa.

As Matt Turck wrote in his latest MAD (Machine Learning, AI & Data) Landscape, we're at the "end of an era", marking a switch from the modern data stack's unbundling to greater consolidation. "In effect, data infrastructure and AI infrastructure are collapsing into one plane; the seams are where value leaks."

## MY PREDICTION

I referred to this report as the "future of the AI-native data stack" because I think we're already in the middle of consolidation, from the separate data and AI infrastructure to a unified stack where data and AI coexist and work together. This stack will be managed by a Data & AI Platform Team, built out of the Data Platform Teams and AI experts.

I think that in the near future, we'll see the merging/evolution of a few current data categories: data lineage into AI lineage, data catalogs and access controls into an AI context layer, workflow orchestration with agentic orchestration, and ETL pipelines with AI model/pipeline orchestration. We'll also see the creation of some new AI-centered categories: agent observability, unstructured data quality/tagging, and evals.

What slowed down innovation in the past was data lock-in, and what will slow down innovation in the future will be metadata and context lock-in. In an increasingly complex world of data, compute, and agents, I think the metadata and governance layer for an enterprise must strategically become the most unified part of the stack. It needs to be open, neutral, and built for collaboration across tools and teams, acting as a unified control plane across an enterprise's diverse data and AI tools without lock-in to any single data source, compute layer, or agent/model.

# Data teams must evolve – they're the best placed people to solve the AI context problem

The rise of AI and unified data & AI stacks aren't just technical shifts — they're organizational too. As infrastructure changes, so do the people powering it.

The AI wave is hitting data teams hard. The data skills that dominated the last decade (writing SQL, building ETL, maintaining dashboards, etc.) are becoming increasingly automated. In the last year, ChatGPT Enterprise has seen an **8x increase** in weekly messages, **19x increase** in structured workflows, and **320x increase** in average reasoning token consumption per organization — mostly used for writing, coding, customer support, and data analysis. This is already starting to replace human execution for early career tasks (think intern and junior roles), according to ADP Research. However, more complex, later stage roles have found success with AI augmentation: for example, a SQL developer called AI "the most powerful tool we've ever seen in our lives". This gap leaves data people confused about their place and relevance in this new AI world.

But here's the paradox: data people are actually the best placed to thrive in this new AI world. Unlike software engineers — who work in deterministic systems where a function either passes or fails, and code is shipped only when it behaves identically every time — data teams have spent their entire careers working in nondeterministic environments. Two analysts can explore the same dataset and reach different conclusions, or two data scientists can follow the same inputs and generate entirely different models. Iteration, ambiguity, and back-and-forth experimentation are the norm among data people.

LLMs behave much more like analysts than like software. They are inherently nondeterministic, and getting a model to reliably answer the same question the same way requires training, semantic guidance, careful prompting, iterative refinement, and lots of exploration. This is second nature for data people, but foreign to deterministic disciplines. Business users often do not experiment with technology, and traditional engineering workflows aren't built around ambiguity. Software engineers ship code if it satisfies the specified goal, checked via a simple "yes/no" unit test.

Data practitioners, on the other hand, are already fluent in exactly the type of reasoning AI requires. Yet despite being uniquely equipped for this new world, many organizations haven't figured out how to position them. In our survey, **31%** of organizations cited lack of AI talent and **23%** unclear ownership of AI as blockers to scaling AI.

## MY PREDICTION

As data and AI converge, and as AI takes on more and more grunt work, old data roles like SQL jockeys and dashboard builders will increasingly become obsolete — but data people will not. Their natural strength in navigating nondeterminism makes them the most valuable people today to take on the challenge ahead: setting AI up for success with semantic training, context engineering, and unified observability.

I think we'll see significant changes in data roles to reflect this. For example, Data Analysts will likely evolve into Analytics Context Engineers, and Data Engineers into Data & AI Engineers (covering orchestration and eval stacks). Data Platform Engineers will end up working more with unstructured data and thus larger scaling problems, including unstructured data tagging, AI lineage, and AI observability.

To learn more about what these new roles look like in the real world, we spoke to top data officers and practitioners from AI-forward companies at this year's Re:Govern conference. Check out distilled highlights about the future of data in an AI world from leaders at Mastercard, Workday, Elastic, and more at the Re:Govern watch center.

# Apache Iceberg will go mainstream and drive the data stack to fundamental interoperability

For years, the data warehouse has been the center of gravity for analytics. Once data lands in the warehouse, everything else follows: compute happens inside the warehouse, and all tools (ETL, BI, governance, active metadata, etc) connect to it. That creates data gravity and vendor lock-in.

Launched in 2019, Apache Iceberg fundamentally changes this — perhaps the biggest change in the data product ecosystem today. An open-source table format that sits directly on top of cloud storage, Iceberg turns stored datasets into high-performance tables with ACID transactions, schema evolution, time-travel queries, and more. This means that regardless of where data is stored, any compute engine can query it with Iceberg, which shifts gravity from compute to storage.

Iceberg even has some interesting financial implications. We've worked with customers whose finance teams changed infrastructure spend from OPEX (operating expenditure) to CAPEX (capital expenditure) after switching to Iceberg. This is because the data, rather than being locked into a single vendor and queried for a single purpose, could now be seen as an asset or IP because of its interoperability across tools.

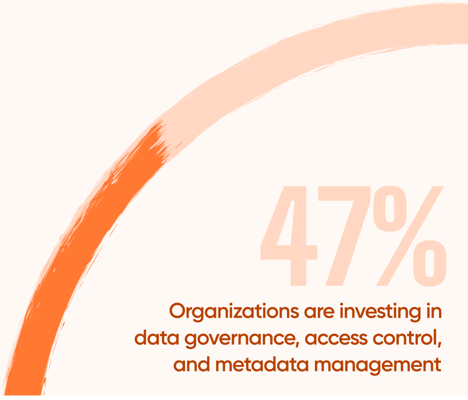
In the last couple of years, Iceberg has become the "industry standard". For example, AWS Athena and Glue, Google BigQuery and BigLake, Snowflake, and Databricks announced native support for Iceberg, including governance through Snowflake Polaris Catalog and Databricks Unity Catalog. Microsoft adopted Iceberg to share data across Snowflake and Fabric, and Salesforce built "zero copy support" for open data lakes with Iceberg.

Meanwhile, dozens of established startups and open-source projects like Cloudera, Dremio, DuckDB, Firebolt, and Trino use Iceberg as their native table format. Apache even hosted the first ever Iceberg Summit in 2025 with widespread industry support.

## MY PREDICTION

As Iceberg becomes the dominant table format, we'll see a push for fundamental interoperability of data. Lock-in will fade and choice will explode as data teams can bring their own compute to any data storage system. This will inevitably lead to the creation of new types of compute engines and business models.

This shift won't be frictionless though. With multiple engines writing to Iceberg tables, we'll see governance, lineage, and context become more important than ever before. That's why it's no surprise that **47%** of organizations said they're now investing in data governance, access control, and metadata management, making it the top investment in our survey.



# Conclusion

If there's one thing this year made clear, it's that we're standing at an inflection point. Despite its challenges, the promises of AI are real — the models are smarter, the tools are more mature, and its business value is unmistakable. And yet the gap between this hype and actual results is also vast.

The predictions I've outlined all point to the same truth: the next era of data won't be unlocked by another feature or framework but instead by rethinking how data, AI, and humans work together. What gives me hope is that this isn't an entirely new challenge. Data teams have spent years navigating ambiguity as technology and best practices constantly evolved under their feet. AI just ups the stakes.

Here's what we're certain about: the organizations that succeed in this new world won't be the ones with the fanciest tools or flashiest models. They'll be the ones that invest in the unglamorous foundations that let AI operate safely and accurately — context, governance, and interoperability.

At the same time, there's still so much we don't know as a community. What will the dominant interface for analytics really look like? Who will ultimately own the AI analyst — the warehouse, the BI tool, the agent platform, or something entirely new? How much autonomy will organizations be willing to give AI systems, and where will humans always need to stay in the loop?

These questions are still unfolding, and the answers will likely surprise us. But if the last decade and even the last year taught us anything, it's that the data world doesn't stand still for long. In this world, being able to constantly rethink from first principles, move fast, and iterate will make the difference between companies that truly transition into an AI world and those that don't. The next chapter is being written now, and I'm excited to see the unprecedented changes it will bring.

atlan