Your vision.
Engineered.

N-iX

N-iX 2027

Data Index

Closing the AI-readiness gap

Executive summary

Most enterprises now have an AI mandate, an AI budget, and a modern data platform. What most lack is the data AI can use. This report measures that gap from inside the estates, where AI ambition meets the data it is supposed to run on.

The constraint on enterprise AI has shifted from the model and the budget to the state of the data and the lack of anyone accountable for it. The data strategies in force today were written for BI and dashboards, for a human reader who tolerates a late pipeline or an ambiguous definition. AI tolerates neither, and it has exposed a decade of deferred decisions about quality, ownership, and documentation.

By AI-ready data, we mean data that is aligned to a named use case, governed at asset level, quality-checked in the pipeline, described by live metadata and lineage, and in scope across structured and unstructured formats.

What we found

Source: N-iX Data and AI Engineering Survey, September 2026, N-iX audience polls, 2025–2026

80%	**Ambition ahead of the data**	Four in five engagements began with the client wanting more from AI than the data could support; 75% had documented a third or less of the problems.
65%	**A modern platform does not close the gap**	Of the 40% of platforms rated modern and AI-ready, about two-thirds still showed the gap; BI-only platforms almost all did.
82%	**The failure is semantic**	Undocumented business logic and conflicting sources of truth are four in five of the top quality failures. Schema drift: 5%.
63%	**Nobody owns it**	Nearly two-thirds of engagements had no structural owner for AI-ready data; 23% of practitioners say nobody keeps AI context current.
14%	**The ready minority planned for it**	One in seven began with ambition matched by data: named owners, written definitions, a narrow first use case, a timeline with preparation.

What it means for leaders

The AI investment with the best return sits upstream of the model: naming owners, agreeing definitions, putting quality gates in pipelines, and bringing the unstructured majority of the estate into scope, sequenced against one narrow use case.

In our engagements, organizations that did this spent their AI budget on AI. The rest funded the same foundation work inside every pilot and recorded it as a delay. This report describes what that upstream work consists of, according to the people who do it.

Table of contents

Authors:



Rostyslav Fedynyshyn
Head of Data & Analytics Practice, N-iX



Danylo Diachenko
Senior Data Architect



Maryana Sydoriv
Market Intelligence Senior Manager

N-iX

Part 01 The gap, quantified

The boardroom mandate meets the data estate 8
What "AI-ready data" actually means 10
Anatomy of the gap: five dimensions 12

Part 02 The Index: evidence from the field

What enterprises ask for vs. what they need 16
What N-iX data engineers find 18
The engineer's view: 258 practitioners on their most recent engagement 23
The experts view: why data strategies fail AI 29
The four enterprise data archetypes 30

Part 03 Closing the gap

The AI-ready data foundation 37
Governance that accelerates instead of blocking 38
Sequencing: the first 90 days vs. the structural work 41
Case studies: closing the gap in practice 43
Closing 45

Introduction

Why this Index, and how we built it

Every enterprise data engagement begins with a request, and almost none of those requests are for AI-ready data. A few weeks in, the same engagement is reconciling two definitions of "customer" or tracing business logic that lives only in an undocumented transformation. This report examines the gap between what was requested and what was done.

Analyst firms and vendors survey thousands of data leaders each year, and their studies converge on one headline: most enterprise data is not ready for AI.[1] They measure what leaders say about their data. This report asks what delivery teams find when they open it.

The N-iX Data Index draws on four evidence streams from delivery, not from surveying leaders. We classified the 185+ enterprise data opportunities that reached our pipeline in the review period by what the customer asked for and what the request implied about the estate behind it.[3] We surveyed 258 N-iX data and AI practitioners in depth about the client estate they most recently worked in.[4] We ran 14 audience polls across our professional channels over two years, collecting 3,104 responses.[5] And we interviewed enterprise CTOs, CIOs, CDOs, and technology partners about why data strategies written before the GenAI era fail when they meet it.

Four evidence streams, all from delivery

185+
enterprise data opportunities, classified by request
Measures what customers ask for

258
N-iX data and AI experts surveyed in depth
Measures what engineers find inside the estate

14 · 3,104
polls and responses
Measures what the market believes

CTO · CIO · CDO
and technology partner interviews
Measures why strategies fail

Where AI ambition meets the data it runs on

A note on rigor. Our evidence base is not a random sample: organizations that bring in a data partner already suspect something is wrong, so our engagements skew toward estates with problems. The practitioner survey is small and deep rather than large and shallow, and we report its sample size with every figure; sub-group figures based on fewer than fifteen engagements are labeled indicative. Some of what our engineers told us runs counter to the convenient narrative, and we kept those findings.

What you will walk away with:

- **A working definition of AI-ready data,** and a clear picture of where the gap sits and why it survives inside organizations that believe their platforms are modern;
- **Field evidence:** what enterprises ask for versus what they need, what engineers find in the first 90 days, and what the ready minority did differently;
- **Four archetypes to locate your organization,** a sequenced playbook, and a self-diagnostic scorecard.

The gap in numbers

Source: N-iX field data. What N-iX engineers found inside the estates, and what 185+ enterprise requests told us

80%

of engagements began with the client wanting more from AI than the data could support

65%

of modern, AI-ready platforms still showed the ambition-to-data gap

75%

of clients had documented a third or less of the critical data problems the engagement found

82%

of the quality failures engineers meet most often are definition failures; schema drift: 5%

63%

of engagements had no structural owner for making data AI-ready

0.54%

enterprise data requests asked for AI-ready data by name

84%

of data and analytics leaders say their data strategies need a complete overhaul before AI can succeed

The market sees the same pattern from outside: 84% of data and analytics leaders say their data strategies need a complete overhaul before their AI ambitions can succeed.[1]

Part I

The gap, quantified

The boardroom mandate meets the data estate

AI has moved from the innovation budget to the revenue plan, and the requests reaching delivery teams have changed with it. A fifth of the 185+ enterprise data opportunities in our pipeline now arrive with an explicit AI or machine learning ambition attached, and many more carry one implicitly: the dashboard request that is really a question about whether the data can be trusted.[3] The wider market shows the same acceleration. In one survey of 2,000 companies, 64% said they had moved advanced AI beyond pilots into production, while 7% had reached the level of data readiness that scaled AI requires.[2]

What rarely arrives is a request for AI-ready data by name. Across the entire pipeline we reviewed, one opportunity asked for it. Customers name a symptom, a container, or an ambition and leave the data's condition unnamed, because, from where they sit, the data looks fine.

Part II puts numbers on the distance between the ambition and the data. The short version: four in five engagements began with the ambition ahead of the data, and the client had usually documented a third or less of the problems that would shape the work.

63% vs 63%: the perception gap, at market scale

Source: Salesforce, State of Data and Analytics, 2025

Business leaders

63%

describe their organization as data-driven

Data and analytics leaders

63%

say their company struggles to drive priorities with data

The closer a person sits to the data, the less ready it looks.

Market research measures the same pattern from the outside. In Salesforce's survey of more than 7,600 business and technology leaders, 63% of business leaders describe their organizations as data-driven. In contrast, an identical share of data and analytics leaders, 63%, say their companies struggle to drive business priorities with data.[1] The same study found that 84% of data and analytics leaders believe their data strategies need a complete overhaul before their AI ambitions can succeed. The two groups look at the same companies and the same data; what differs is how close they sit to it. The closer a person is to the data, the less ready it looks, and our engineers work at that closest point.

For organizations that pushed AI into production anyway, the cost is measurable. Among data leaders whose companies run AI in production, 89% report having seen inaccurate or misleading AI outputs, and 42% lack full confidence in the accuracy and relevance of what their AI produces.[1] More than 80% of enterprises now delay, limit, or alter AI initiatives at least occasionally because of data-related risks.[2] Analysts expect that through 2026, organizations will abandon 60% of AI projects that lack AI-ready data.[6]

89%

of data leaders with AI in production have seen inaccurate or misleading AI outputs

Salesforce

42%

lack full confidence in the accuracy and relevance of their AI outputs

Salesforce

80%+

of enterprises delay, limit, or alter AI initiatives because of data-related risks

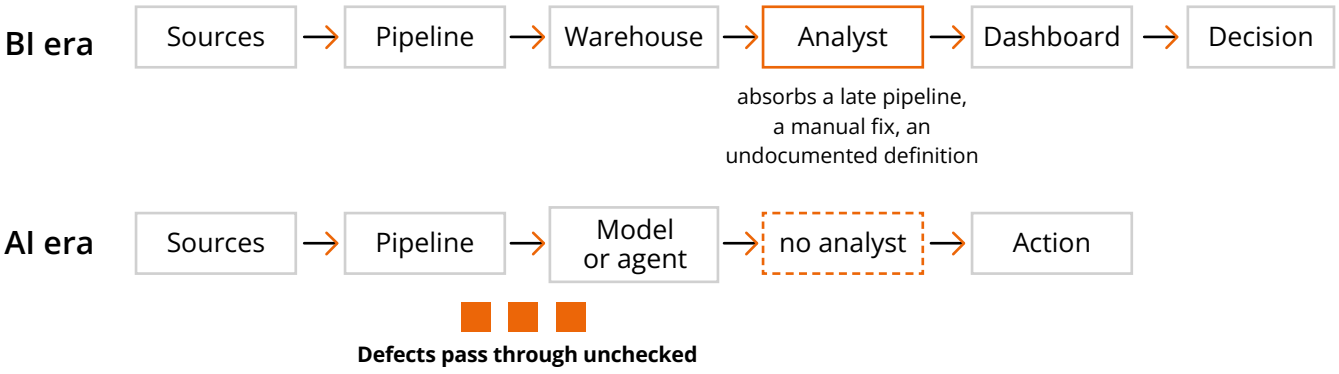
Salesforce

60%

of AI projects without AI-ready data will be abandoned through 2026

Salesforce

Why the bill has come due now



AI has repriced a gap that was already there.

Most enterprises spent a decade investing in data warehouses, lakes, BI platforms, and dashboard layers, and that investment shows in our survey, where 40% of platforms are modern. The problem is what those investments were meant to do. They summarize the business for human decision-makers, on a reporting cadence, from structured tables. A quarterly dashboard is forgiving: it tolerates a late pipeline, a manual correction, or an undocumented definition, because an analyst absorbs those defects before anyone else sees them. Models and agents consume data directly, continuously, and at scale, and every deferred decision about quality, ownership, and documentation that a dashboard absorbs is now exposed in production. AI has repriced an existing gap. The undocumented business logic that once cost an analyst an afternoon now costs an AI initiative its credibility.

What “AI-ready data” actually means

When our engineers say a client’s data is not ready for AI, they rarely mean it is dirty in the conventional sense. The failures they meet most often are failures of meaning: business logic buried in transformations that nobody documented, and two sources of truth with defenders on each side (Finding 4). The data is present and technically valid, and nobody can say with confidence what it means or which version is right.

That is why “our data isn’t ready for AI” has become a phrase everyone repeats and few can act on. The published research shows the same confusion at scale: 63% of organizations either lack the right data management practices for AI or are unsure whether they have them.[2]

The working definition

AI-ready data meets five conditions

Use-case-aligned	Represents the specific use case it serves, edge cases and outliers included
Governed at the asset level	Ownership, permitted uses, and sensitivity defined per asset, before the pipeline is built
Continuously quality-assured	Automated checks in the pipeline, at the cadence the model consumes data
Described by live metadata and lineage	The data carries where it came from, what transformed it, and what it means
In scope across formats	Documents, communications, images, and logs alongside the structured tables

Undocumented business logic is a metadata failure, and conflicting sources of truth are a governance failure; both are invisible to a quality check that tests whether values are present and well-formed. The fifth condition is the one most estates have not started on.

BI-ready is not AI-ready

Source: HBR Analytic Services / Hyland

	BI-ready: built for an analyst	AI-ready: built for a model or agent
Consumer	An analyst	A model or agent
Cadence	Reporting cycle, batch	Continuous, at the speed of the action
Scope	Curated structured tables	The full estate: documents, tickets, logs
What a defect does	A wrong number someone questions	A wrong output nobody sees, or an action
Missing definition	An analyst compensates	The model answers fluently and wrongly
Readiness is	Platform deployed, dashboards live	Stated conditions met for a named use case

65% say structured data is prepared for AI

39% say the same of unstructured data

The most expensive misunderstanding in enterprise data is treating a mature BI estate as evidence of AI readiness. A BI estate was built for an analyst who absorbs defects on a reporting cadence from structured tables; an AI system consumes the whole estate continuously, with no human between a data defect and a wrong output, and most enterprise knowledge sits in the contracts, emails, tickets, and code that a decade of data investment left out of scope.[7] On platforms that were modern but built for BI only, the ambition-to-data gap was almost universal (Finding 1). AI has repriced a gap that was already there

Sidebar: what agentic AI adds to the bar

Every condition above assumes AI that produces an output a human reviews. Agents remove that assumption: they retrieve data across systems, chain decisions, and initiate actions. That raises the definition in three ways. Data must carry semantic context, the shared business meaning of terms and entities. It must be refreshed at the speed of action, not the speed of the report. And someone must own keeping it that way; section 15 takes this on. The practical consequence is a shift in the question. “Is our data AI-ready?” cannot be answered in general. “Is this data ready for this use case?” can.

Anatomy of the gap: five dimensions

"Our data is not ready for AI" is true in most enterprises and actionable in almost none. Readiness fails in specific, diagnosable ways, and across our delivery work the failures cluster into five dimensions. They interact: a governance failure surfaces as a quality incident, and a 2019 architecture decision becomes an access constraint today. Still, different people own them, and different work closes them. Each is summarized below with one field figure and one market figure; the evidence is in Part II.

Five dimensions of the gap

01 **Quality and consistency**	**N-iX field** 26% name it the biggest first-90-day blocker; 44% the reason AI stalls	**Market** 89% of leaders with AI in production have seen bad outputs	**Owner:** Domains that produce the data
02 **Access and architecture**	**N-iX field** Access approvals block 16% of engagements; legacy architecture 7%	**Market** 19% of data siloed or unusable; 70% think their best insights sit there	**Owner:** Platform and architecture
03 **Governance and trust**	**N-iX field** 63% of engagements had no owner for AI readiness	**Market** 43% of data leaders have a formal governance framework	**Owner:** Central office for standards; domains for content
04 **Metadata and context**	**N-iX field** 2% call it the biggest blocker, yet 54% hit undocumented logic most	**Market** 6% have a unified logical view of their data	**Owner:** Platform for tooling; domains for definitions
05 **Talent and operating model**	**N-iX field** 58% of successful projects had named data owners; 28% reliable pipelines	**Market** 43% name data readiness among their biggest obstacles	**Owner:** The executive sponsor

Dimension 1:

Quality and consistency. 26% name it their biggest first-90-day blocker; 44% name it the reason AI stalls

The failures behind the word are rarely the ones tooling catches: two systems reporting different values for the same fact, business logic in transformation code no one can explain, records incomplete in ways an analyst would notice and a model will not. Schema drift, which observability tools are built to detect, was the least cited of the four (5%). Most quality problems in enterprise data are definition problems, and AI has raised the price: a dashboard shows a defect as a wrong number someone questions, while a model trains on it.

Owner: the domains that produce the data.

Dimension 2:

Access and architecture. Access approvals block twice as many engagements as legacy architecture (16% vs 7%)

Data that exists but cannot be reached, joined, or served at the speed a use case demands. In our survey, this is friction, not legacy technology: only one respondent walked into a mostly legacy, on-premises estate. The modernization wave has happened, and it left fragmentation behind. The average enterprise runs close to 900 applications, fewer than a third connected, and data leaders estimate that 19% of their company's data is siloed or unusable. In contrast, 70% believe their most valuable insights sit in that share.[1] Nearly 80% of enterprises say data access challenges are holding their AI initiatives back.[8]

Owner: platform and architecture.

Dimension 3:

Governance and trust. 63% of engagements had no owner for AI readiness

Governance in most enterprises was designed as a control function: approve access, restrict sensitive fields, pass audits. AI puts it in a double bind we call the trust paradox. Without governance, AI is unsafe to deploy, since no one can say which data may feed which model. With governance built for the BI era, AI is too slow to deploy, since every use case queues behind a review process designed for quarterly reporting changes.[1] Underneath the bind is an ownership vacuum: in nearly two-thirds of the engagements our engineers describe, no one structurally owns making data AI-ready, so the work falls to the AI project team ad hoc, or to no one.

Owner: a central data office for standards, domains for content

Dimension 4:

Metadata and context. 2% call it their biggest blocker; 54% say undocumented logic is the failure they hit most

Almost none of our engineers call missing metadata their biggest early blocker. Yet the failure they report most often, undocumented business logic, is a metadata failure by another name, and documented business definitions top their list of fixes. An analyst compensates for missing context; a model knows only what the metadata tells it, and without lineage and definitions it produces fluent output built on the wrong data. Of the engineers with GenAI experience, 61% say the hardest part of unstructured data is evaluation: no reliable way to tell whether retrieval is good enough.

Owner: data platform for the tooling, domains for the definitions.

Dimension 5:

Talent and operating model. 58% of successful projects had named data owners; 28% had reliable pipelines

In our experience, this dimension decides whether the other four get fixed. Data teams were staffed and measured to keep pipelines running and serve the BI backlog. AI readiness adds a workload closer to product engineering, and few organizations have re-planned capacity or assigned ownership, so remediation happens ad hoc inside AI project budgets, invisible to leadership. The market sees the same thing: 43% of data leaders name data readiness as among their biggest obstacles to AI.[10] On the projects that went well, named data owners and documented definitions were the two conditions engineers most often found in place; reliable pipelines, the only purely technical factor, ranked well behind (Finding 8).

Owner: the executive sponsor.



Reading the dimensions together

Two observations follow. Enterprises rarely fail on one dimension; an estate with poor metadata almost always shows quality incidents, because no one can see what changed, and our survey shows the same failure named under three different headings. The dimensions also differ in remediation horizon. Quality and metadata respond to focused engineering within a quarter, while architecture and operating model take several. That asymmetry is why a readiness profile is more useful than a single readiness score.

Surya Singh,

Founder, BoxNCase

The clearest tell is that the strategy defines dashboard metrics but not stable entity identity, provenance, freshness, or retrieval evaluation. BI can tolerate collapsing records after the fact. An AI system cannot. In specialty-food commerce, a dashboard may only need a category total; retrieval must distinguish the manufacturer, product identity, pack and case configuration, handling constraints, current orderability, and when each fact was last verified. Otherwise the model can produce a fluent answer about the wrong product or an outdated offer.
If an RFP does not assign ownership for source normalization, freshness rules, retrieval-quality tests, and failure handling, the data-readiness work has not really been scoped. It has been deferred until the model appears to fail.

Part II

The Index: evidence from the field

What enterprises ask for vs. what they need

A sales pipeline is an unusual research instrument, but it records what organizations were willing to pay for, in their own words, before anyone from outside had examined their data. We classified the 185+ enterprise data opportunities that reached N-iX in the review period by the demand pattern in the customer's description, then compared those patterns with the work the engagements required.[3]

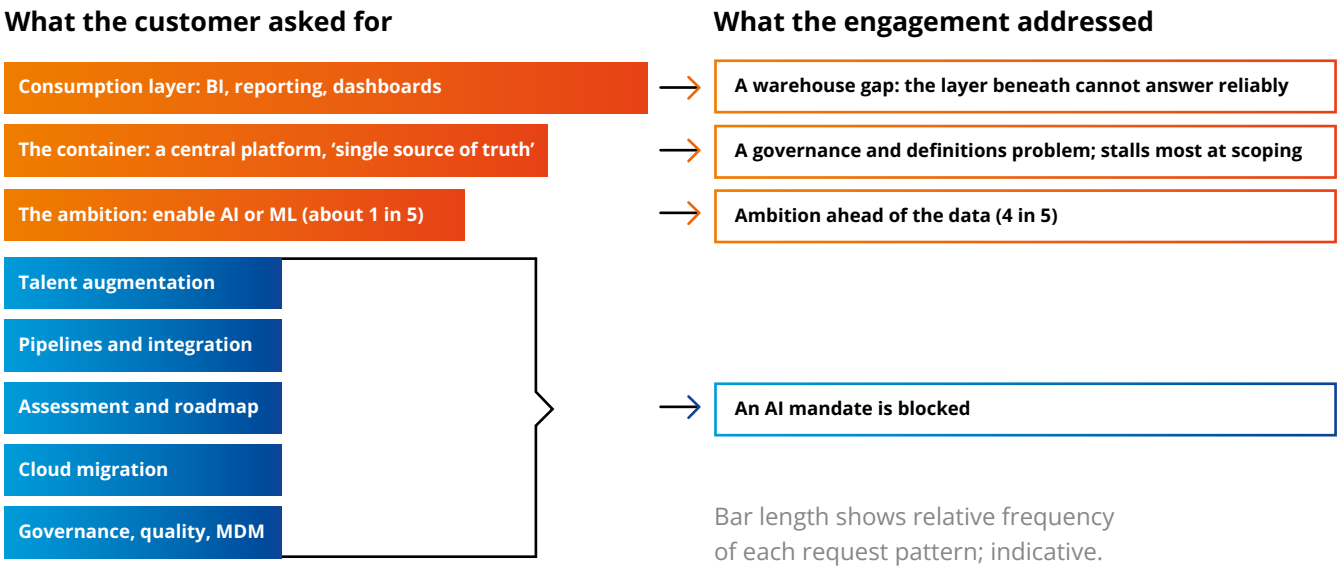
What they ask for: dashboards, a platform, and, in one request in five, AI

The most common request is the consumption layer: BI, reporting, dashboards, often a named visualization tool. Close behind is the container: build or modernize a central data platform, almost always framed as the wish for a single source of truth. Third is the ambition: enable AI or machine learning on the data, a request that now accompanies roughly one opportunity in five. Behind these come talent augmentation, pipelines and integration, assessment and roadmap work, cloud migration, and governance, quality, and master data management.

Two absences stand out as well. Data readiness for AI, as a request in its own right, appeared in a single opportunity across the set. Data and cloud sovereignty, an active service line and a live boardroom topic, did not reach the opportunity stage once. Every one of these requests assumes the data underneath is already fit for AI.

What they ask for vs. what they need

Source: N-iX analysis of 185+ enterprise data opportunities



What they need: the layer underneath the request

The engagements tell a different story. The dashboard request is frequently the visible symptom of a warehouse gap: the client wants a consumption layer because the layer beneath it cannot answer the question reliably. The platform request comes from scattered data and a wish for a single truth, which is a governance and definitions problem as much as an engineering one; it is also the demand pattern that most often stalls between request and commitment, because scoping is where the customer discovers how much of the work is organizational. Governance requests reach us when governance is already blocking something bigger, and in this review period that something was usually an AI mandate.

What this means for how AI is bought

If an enterprise's request language leaves out data readiness, it budgets and schedules AI programs without the largest component of their actual work. The groundwork migrates into the AI project, unplanned and unfunded, where it becomes the delay and the overrun that later research attributes to AI's disappointing returns. A CTO reading this should ask which line item of their AI roadmap covers the work our engineers spend their first 90 days on. In the roadmaps we see, it is rarely there.



What N-iX data engineers find

The practitioner signal

Before turning to what our engineers find inside client estates, it is worth establishing what practitioners at large believe. Between September 2025 and August 2026, we ran 14 audience polls across our channels and collected 3,104 responses from data engineering, analytics, AI/ML, and technology leadership practitioners. Answer: **enterprise AI is constrained by the state of the data and the capabilities of the people around it.**

What 3,104 practitioners say is holding AI back

Source: N-iX audience polls, September 2025 – August 2026

Signal 1. Where the time goes

46%
60 to 80%
Share of AI lifecycle on data prep

50%
Organizing data
Which part eats the time

48%
During data prep
When bias is cheapest to fix

Signal 2. What blocks scale

47%
Limited data quality
#1 data gap when scaling GenAI

48%
Data literacy
Biggest GenAI adoption challenge

62%
Finding skilled professionals
Challenge implementing new tech

#1
AI literacy and skills
Barrier to AI strategy
(Reddit, ranked)

Signal 3. Diagnosis and ownership

46%
Model's limited knowledge
Main cause of hallucinations

38%
Dedicated function
Who keeps AI context current

23%
Nobody yet
Who keeps AI context current

Signal 4. Trust and ambition

85%
Data security risks
Barrier to enterprise agents

64%
Cybersecurity and privacy
Open-source AI challenge

48%
Multi-agent collaboration
Most value in three years

47%
Predictive analytics
Most practical agent value

52%
Faster dev, same release
AI coding impact so far

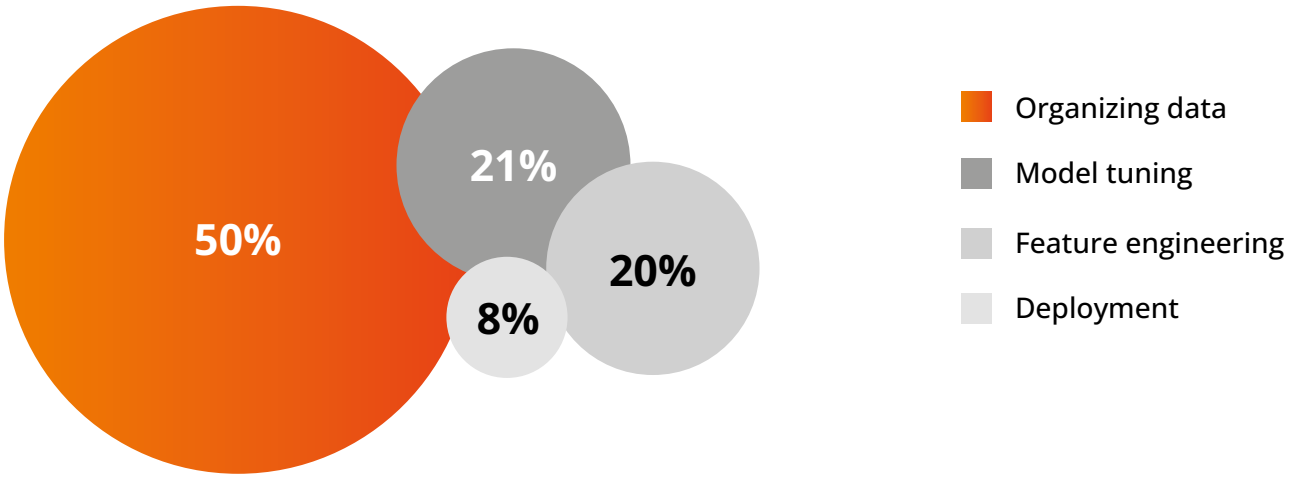
Signal 1: Practitioners believe the 60-80% rule; our engineers measure less

Nearly half of respondents put data preparation at 60 to 80% of the total AI lifecycle, and half of that time goes to organizing data. Our engineer survey shows what that organizing consists of: reconstructing undocumented definitions and reconciling conflicting sources. If preparation dominates the lifecycle and upstream fixes are the cheapest, the AI line item with the best return is the work done before any model touches the data. Our engineers put the figure lower, which we read as the difference between the whole AI lifecycle the poll asked about and a single client engagement the survey asked about.

Where the AI lifecycle goes

Source: N-iX audience polls September 2025–August 2026

Which part eats the time



Share of the AI lifecycle spent on data preparation

52% put it at 60% or above



N-iX engineers put preparation at 30%+ of effort in 70% of engagements (Finding 3).

Signal 2: What blocks scale is data condition

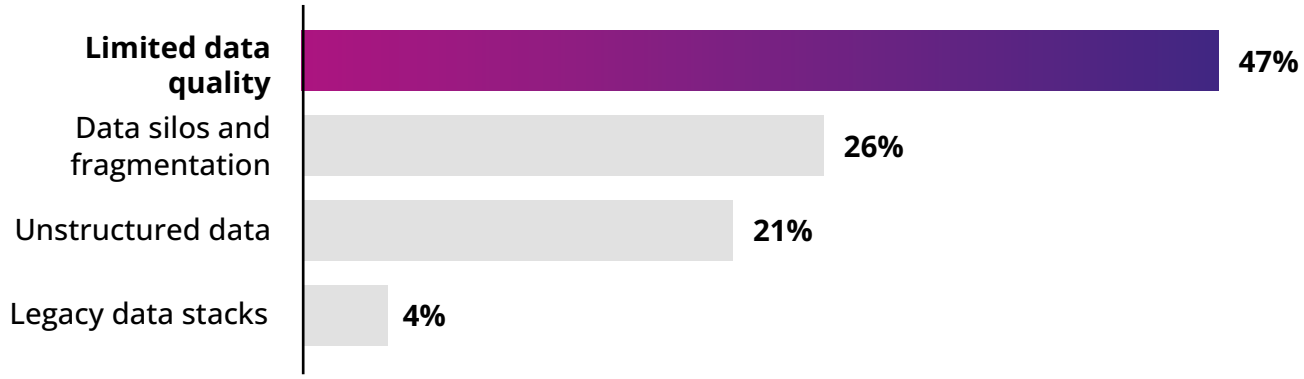
On the number-one data gap when scaling GenAI, 94% of respondents chose data condition, and 4% chose legacy technology; funding ranked last on every channel that asked, behind data literacy and skills. Practitioners are describing a readiness problem, split between the data and the people expected to work with it.

What blocks GenAI at scale

Source: N-iX audience polls September 2025–August 2026

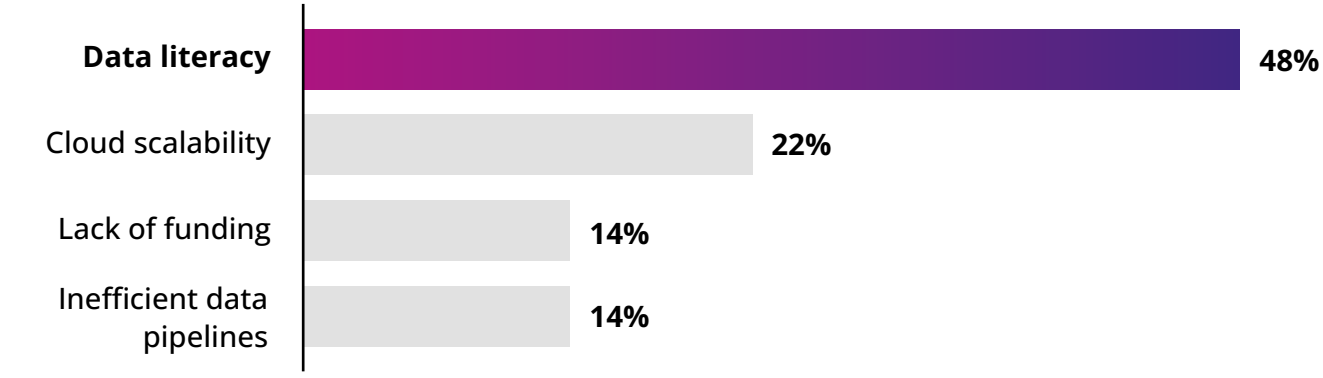
The #1 data gap when scaling GenAI

94% name a data-condition gap



The biggest challenge to GenAI adoption

62% name finding skilled professionals as the top challenge with new tech



Signal 3: A diagnostic gap, and an ownership vacuum

The most consequential finding is a contradiction. The same audience that ranks data quality as the top scaling blocker attributes hallucination, the most visible AI failure, mainly to the model's limited knowledge, and only one in seven to poorly managed data. Practitioners see the damage and locate its origin downstream of where their own answers place the cause. An organization that attributes hallucinations to model limitations buys a better model; one that attributes them to data condition and stale context fixes its retrieval layer and its metadata. Only the second spend addresses what the field, in every other framing, identifies as the root cause. The gap has an organizational counterpart. Who keeps AI's context accurate and current? Most say a data team or a dedicated function; nearly a quarter say nobody yet.
The signal matches delivery: context maintenance is emerging as a discipline faster than organizations are assigning anyone to it.

The diagnostic gap, and who owns context

Source: N-iX audience polls September 2025–August 2026

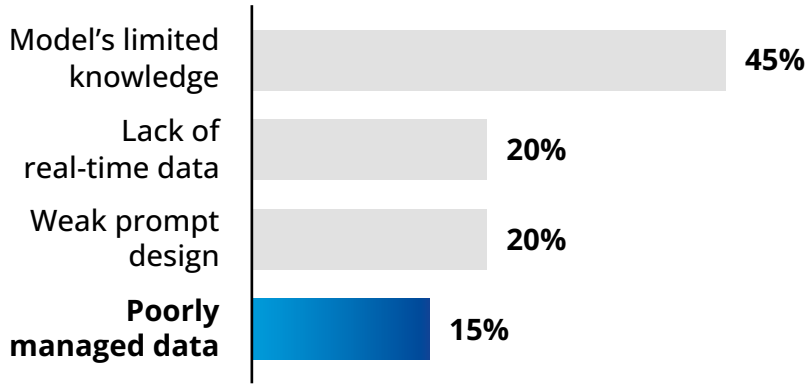
#1 scaling blocker

47%

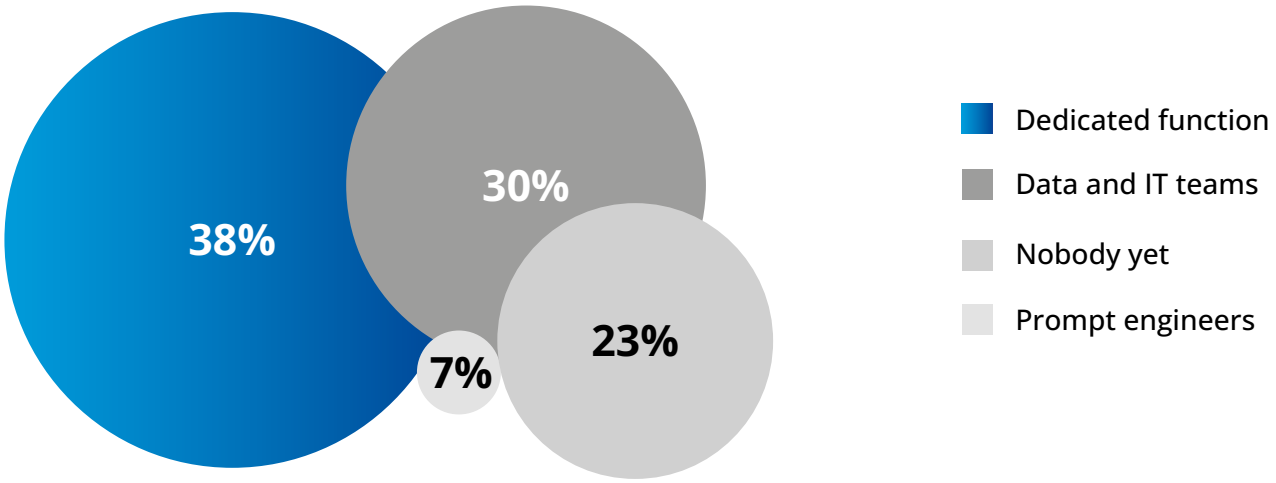
limited data quality

same audience →

Main cause of hallucinations



Who keeps AI context current?



Signal 4: Trust is the brake, ambition is the accelerator

Data security risk dominates practitioners' view of what blocks enterprise AI agents, while nearly half expect multi-agent collaboration to deliver the most enterprise value over the next three years. The same practitioners who report that preparation consumes most of the lifecycle and that nobody owns context expect the next wave of value from the most context-hungry architectures available. Read chronologically, the polls trace attention moving upstream, from skills and security in the earliest questions, through data condition in the middle of the period, to context ownership and agents in the most recent. The market is asking better questions than it was a year ago.

Trust is the brake, ambition is the accelerator

Source: N-iX audience polls September 2025–August 2026

Brake

85% name data security risk the biggest barrier to enterprise AI agents

64% put cybersecurity and privacy first for open-source AI

Accelerator

48% expect multi-agent collaboration to deliver the most enterprise value in three years

47% see predictive analytics as agents' most practical value

Where poll attention moved, September 2025 → August 2026

Skills and security → Data condition → Context ownership and agents



The engineer's view: 258 practitioners on their most recent engagement

Audience polls tell us what the market believes; our own engineers tell us what is there. In September 2026, we surveyed 258 N-iX data and AI practitioners about their most recent client engagement: the platform they walked into, the blockers they hit in the first 90 days, where their time went, and what separated the projects that succeeded. It is the only dataset in this report drawn from people who have seen the data first-hand.

Who answered, and the Index in three numbers

Source: N-iX Data and AI Engineering Survey, September 2026

60%
more than six years in technology

47% · 23%
data engineers · ML or AI engineers

58%
of clients above 5,000 employees

35% · 19% · 14% · 9%
healthcare · retail · logistics · finance

The Data Index 2027 in three numbers

80% ambition ahead of the data

63% no structural owner for AI readiness

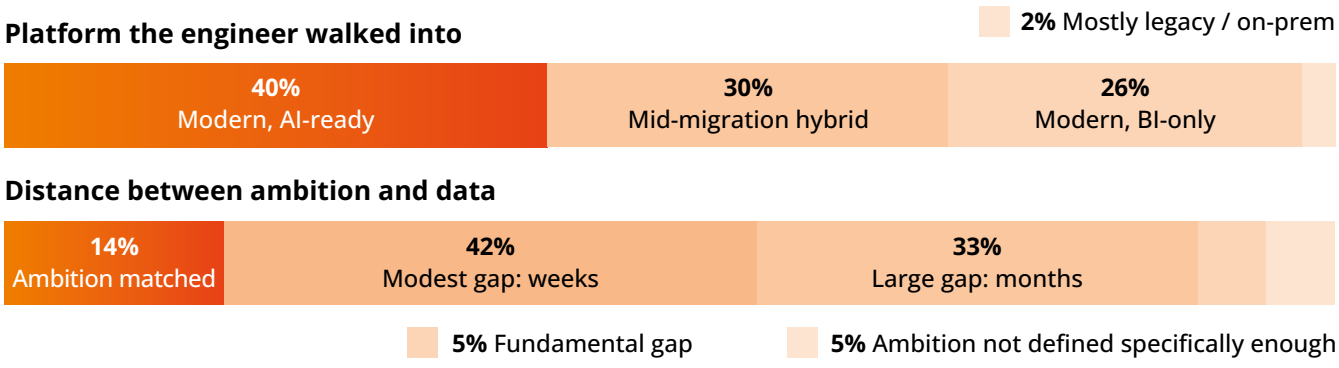
75% clients had documented a third or less of their problems

Finding 1: 65% of modern, AI-ready platforms still showed the gap

Modernity did not close the gap. Two-fifths of engineers walked into a platform they describe as modern and AI-ready, yet only one in seven saw client ambition matched by the data; on modern, AI-ready platforms, about two-thirds of engagements still showed a gap, and on modern, BI-only platforms, all but one did.

Platform on arrival vs. size of the gap

Source: N-iX Data and AI Engineering Survey, September 2026



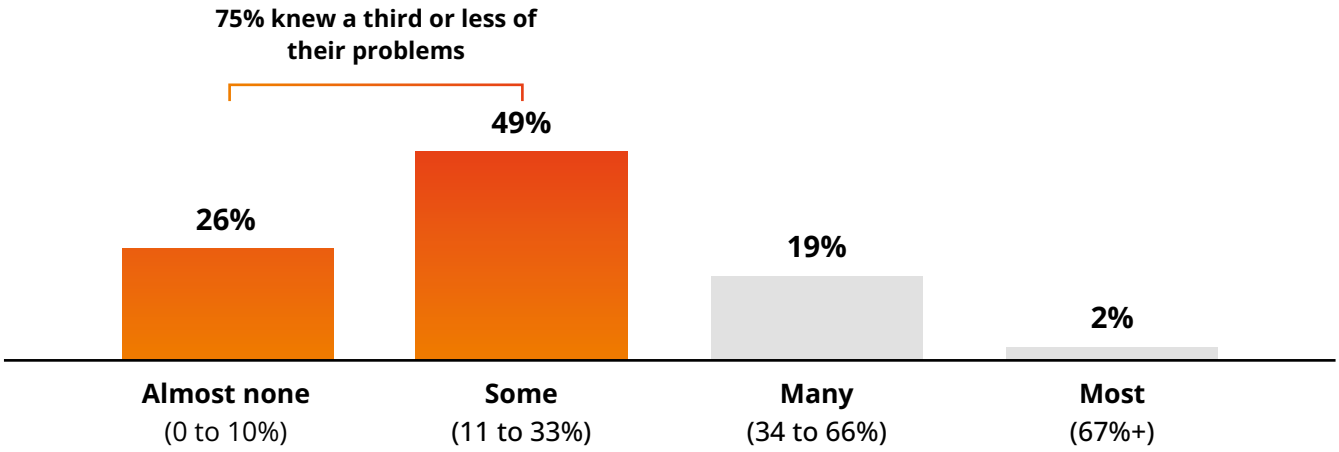
Finding 2: 75% of clients had documented a third or less of their critical data problems

Half of our engineers found that the client had documented between a tenth and a third of the critical data problems in advance, and a quarter found almost none. Three quarters of engagements began with the client aware of at most a third of the problems that would shape the work. This is the perception gap measured from the inside, and the single finding that best explains why AI programs are scoped without their largest component.

How much the client had documented

Source: N-iX Data and AI Engineering Survey, 2026

Share of critical data problems documented before the engagement



Jack Wright,

Senior Vice President of Operations, HealthWright Technologies

When one of ours agents underdelivers, it is almost never the model. It is the data, and specifically which source the system was told to trust. Our clearest case: an automated compliance check used a summary document as its reference instead of the system of record. The summary was out of date, so the check did not just miss errors, it approved them, two runs in a row, and our content tools kept repeating them. When we finally swept everything at once instead of reviewing small batches, the same error turned up 65 times across 40 pieces of content. My stance: measure where each answer came from, not just whether it reads well. A reference that only agrees with the thing it checks proves consistency, never truth.

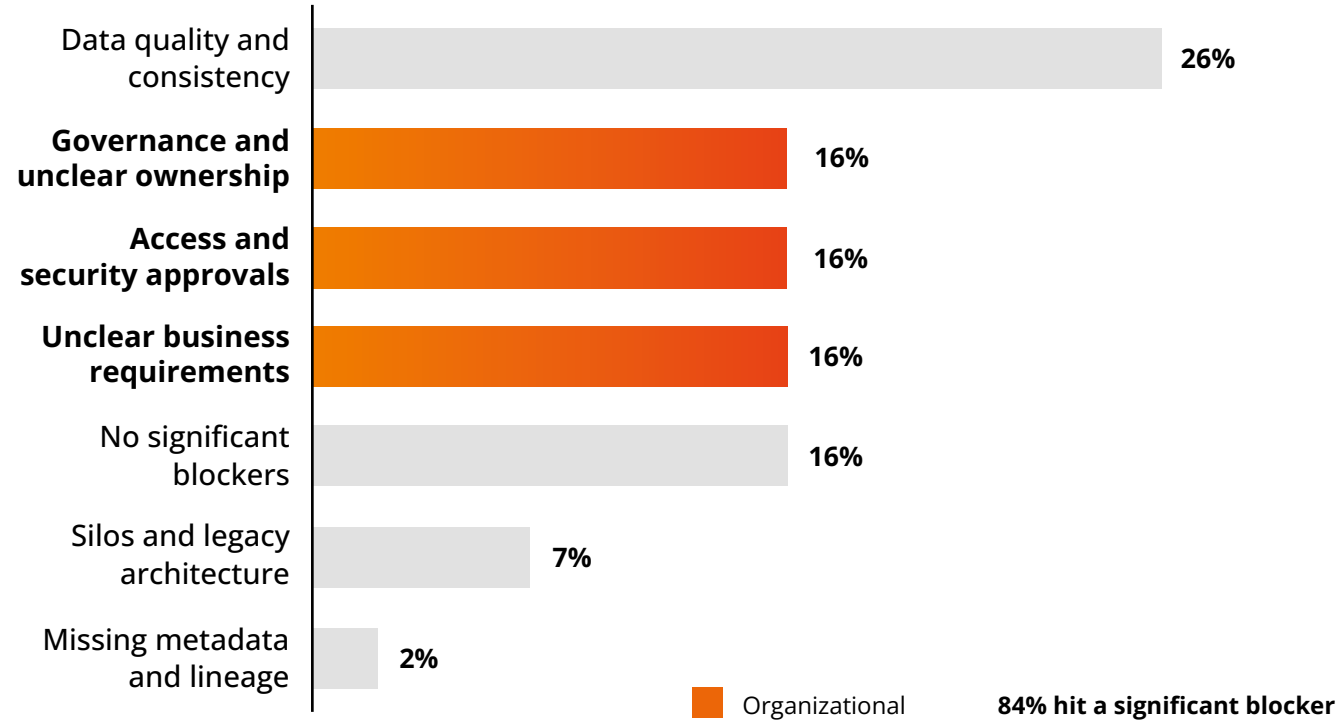
Finding 3: 84% hit a significant blocker in the first 90 days, and most blockers were organizational

The single biggest early blocker splits widely, with data quality in the lead and governance, access approvals, and unclear requirements tied behind it; 84% hit a significant blocker in the first quarter, and the blockers are mostly organizational. Foundation work took under a quarter of the engagement for about half of engineers and more than half of it for about a quarter. These figures are lower than the 60-80% practitioners report in our polls (Signal 1); the two measure different things, and neither supports treating preparation as a rounding error.

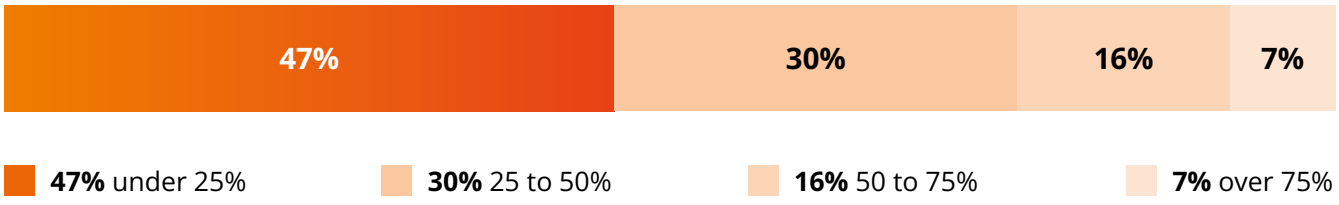
First-90-day blockers and the foundation burden

Source: N-iX Data and AI Engineering Survey, September 2026

Single biggest blocker in the first 90 days



Share of time fixing foundations



Data preparation took 30%+ of effort in 70% of engagements.

Finding 4: 82% of the most common quality failures are definition failures; schema drift is 5%

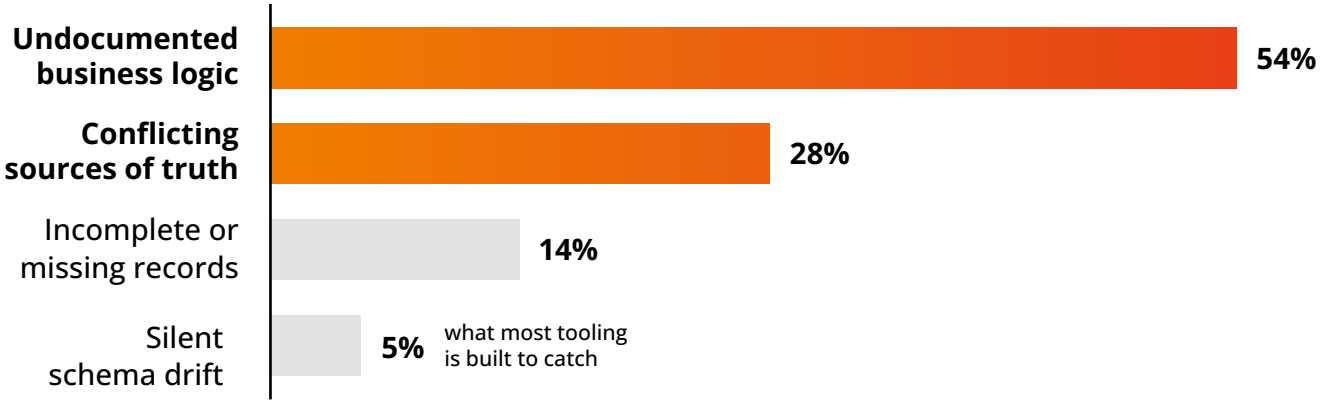
Career-wide, the quality failure engineers hit most often is undocumented business logic buried in transformations, then conflicting sources of truth; silent schema drift is last. Given one fix before data reaches a pipeline, the largest group would choose consistent, documented business definitions. What breaks most often is meaning, and what engineers most want fixed is meaning. Quality tooling built to catch drift and nulls is aimed at the smaller half of the problem.

The failure is semantic

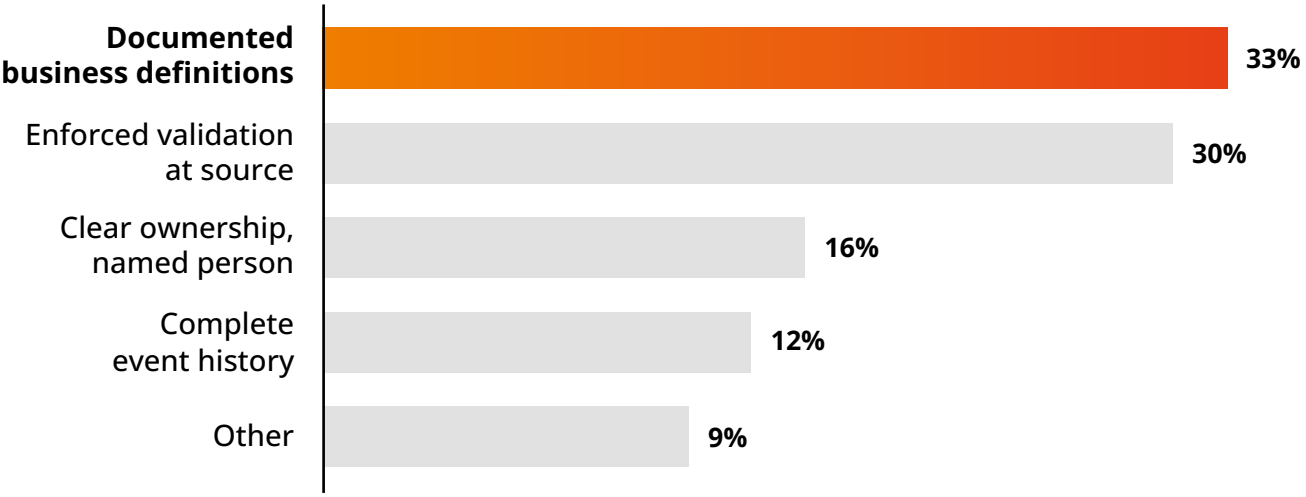
Source: N-iX Data and AI Engineering Survey, September 2026

Quality failure hit most often, career-wide

82% are definition failures



One thing to fix before data reaches the pipeline



Finding 5: 61% of GenAI engineers have no reliable way to judge retrieval quality

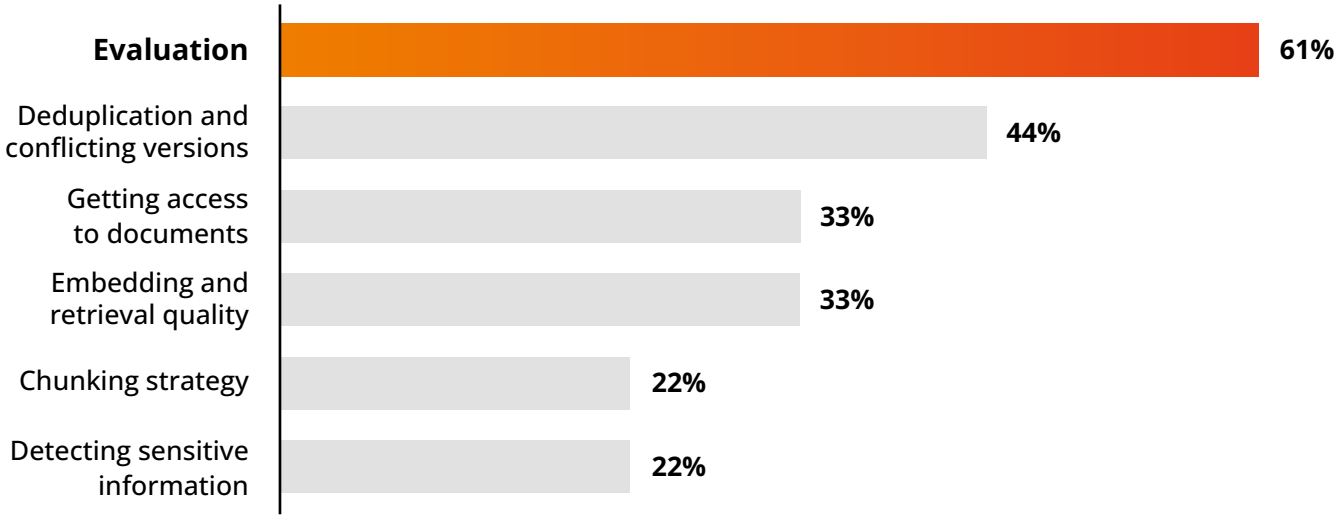
Two in five of our engineers have prepared unstructured data for a GenAI, RAG, or LLM project, and for them the hardest part is evaluation: most have no reliable way to tell whether retrieval is good enough. An organization can chunk and embed its documents in weeks. Knowing whether the result is good enough to act on is the unsolved part.

The unstructured frontier

Source: N-iX Data and AI Engineering Survey, 2026

Hardest part of making unstructured data usable

42% have prepared unstructured data for GenAI

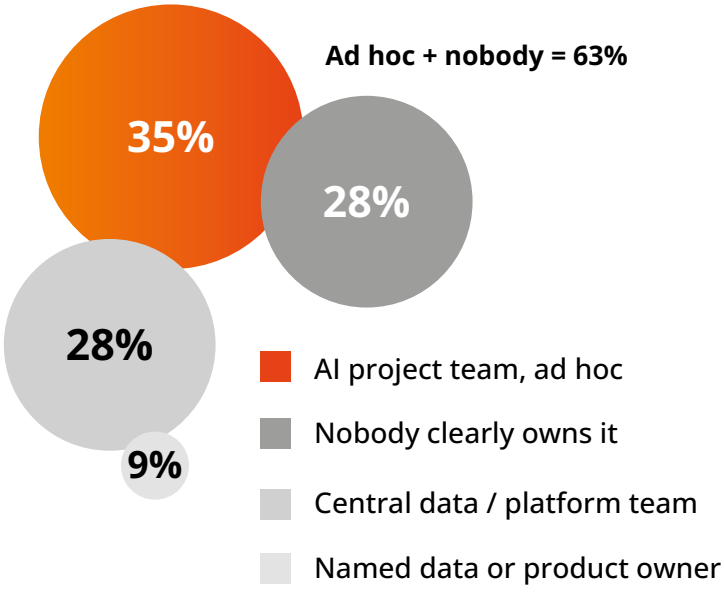


Finding 6: 63% of engagements had no structural owner for AI-ready data

Ownership of AI readiness at the typical client sits with the AI project team, ad hoc, or with nobody; a named data or product owner is rare. Sixty-three percent of engagements begin with no structural owner for the work this report is about.

Who owns AI readiness

Source: N-iX Data and AI Engineering Survey, 2026



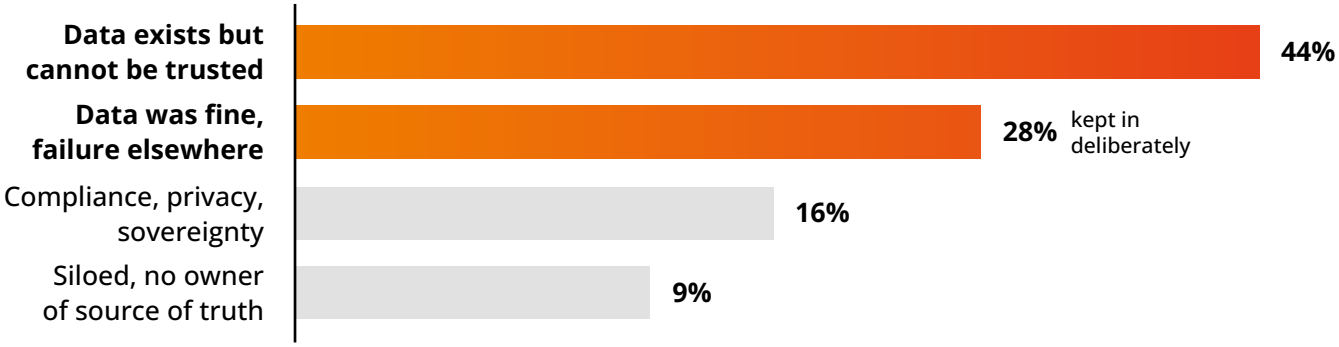
Finding 7: 44% of stalls: the data exists but cannot be trusted. 28%: the data was fine

The most common data-related reason an AI initiative stalls is that the data exists but its quality is too poor to trust. And 28% gave an answer that cuts against this report's premise: the data was fine, and the failure lay in scope, budget, or business definitions. We include that figure deliberately. Data readiness is one of several foundations, and "business definitions," which appears in the "elsewhere" answer, is arguably a data problem under another name.

Why AI initiatives stall

Source: N-iX Data and AI Engineering Survey, 2026

Most common data-related reason an AI initiative stalls

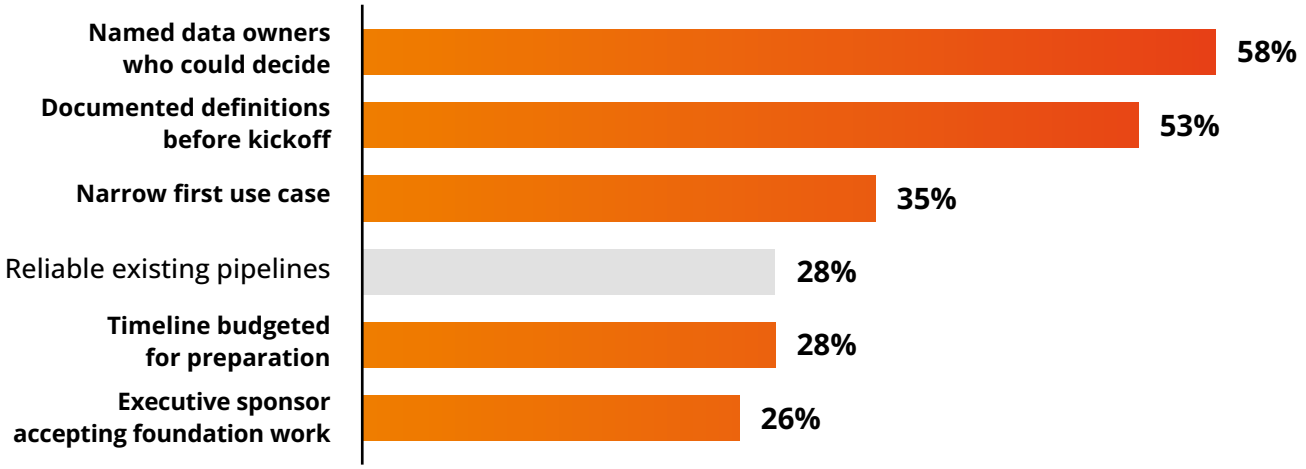


Finding 8: 58% of successful projects had named data owners; 53% had definitions before kickoff

On the projects that went well, engineers most often found named data owners who could make decisions and data access with documented definitions ready before kickoff; a narrow first use case, a timeline that budgeted for preparation, an executive sponsor, and reliable pipelines followed. One of the six is a property of technology. The rest are decisions, and every organization in our sample had access to them before the first engineer arrived.

What the projects that worked had in common

Source: N-iX Data and AI Engineering Survey, 2026



The experts view: why data strategies fail AI

Our engineers see the gap from inside the estate. To test their findings against the buyer's side, we asked data and AI leaders in financial services, pharmaceuticals, manufacturing, and commerce why data strategies built for reporting break when AI arrives.

Their answers fall into 3 failures categories:

- The strategy tests for the wrong consumer
- No common agreement on what an entity is
- Data ownership after AI-based product release

The strategy tests for the wrong consumer

A BI strategy counts as successful when the dashboard reconciles. The leaders we spoke to, including Farhat Siddiqui, Head of Generative AI Enablement & Transformation at Genmab, treat that as the clearest sign a strategy was not written for AI, because a reconciled report says nothing about whether a model can retrieve the right evidence, for the right user, at the right time.

Experts proposed a simple test for a document-based assistant, answer an ordinary policy question from the current approved document, keep its exceptions, cite the source, and respect the user's access rights. Most reporting strategies cannot pass it. They never specified which document is authoritative, how versions are handled, who may see what, or who corrects a wrong answer.

Structured data has the same gap. Reporting needs the current state of the data, but models need to know what the data looked like at the moment of a past decision, with lineage at the level of individual features. Venkat Navneeth Burla, Senior Data Scientist at Vertex Pharmaceuticals, described a propensity model that validated unusually well because a feature built from two vendor datasets with different timestamp logic carried information that would not have existed at prediction time. The problem came from the historical join. His team caught it before deployment and moved the time-based validation split into the pipeline.

Nobody agreed on what the entity is

Experts named entity resolution as the foundation they most often find missing. When data comes from several upstream sources, the same customer, product, or account usually has different identifiers, timestamp formats, and validity rules in each one. Every model and agent built on top inherits that ambiguity, and the fix gets more expensive the further downstream it happens.

Quality is treated as a project, not a responsibility

The leaders described data problems that were known, documented, and never fixed, because the team that saw the damage had no mandate to change the source systems, and the team that owned those systems could not see the damage. In one deployment, an AI classifier launched at around 60% accuracy and passed 90% without any change to the model. The only change was giving a business team ownership of the correction loop. The organizations that had moved furthest ran critical datasets as products, with named owners, versioned contracts, and quality commitments, so that a broken schema became the owner's problem.



The common root

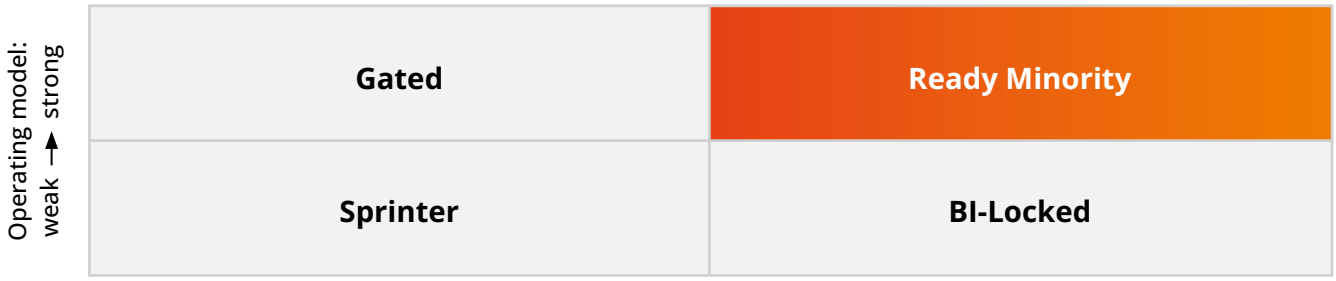
Experts describe the same pattern our engineers measure in Finding 6. Data-readiness work is left out of the scope: no owner for source normalization, no freshness rules, no tests of retrieval quality, and no plan for failures. The gap is unknown until a model starts producing wrong outputs.

The four enterprise data archetypes

Readiness gaps cluster into recognizable organizational types, each with a presenting symptom, a hidden cost, and a first move. We describe four. They come from delivery experience, and the engineer survey lets us size them approximately; the survey did not ask engineers to classify their client, so the proxies overlap. For each, we also name a public N-iX engagement that started in that state.

The four enterprise data archetypes

Source: N-iX Data and AI Engineering Survey, 2026



Sprinter proxy 38%	BI-Locked proxy 26%	Gated proxy ~1/3	Ready Minority proxy 14%
AI-ambitious, data-poor; pilots everywhere, no owner	A modern estate for the wrong consumer	Governance without flow; use cases queue behind reviews	Ambition matched by data; foundation work was planned
Hidden cost The same preparation done several times	**Hidden cost** A second, unplanned modernization	**Hidden cost** Caution compounds	**Hidden cost** None; data has stopped being the constraint
First move **Name an owner, pick one use case, pause parallel pilots**	**First move** **Audit against the AI-ready definition, not the BI one**	**First move** **Governance per use case; gates to guardrails**	**First move** **Institutionalize and move to a portfolio**
In the field: UK financial provider	In the field: E-commerce and subscription platform	In the field: Global managed cloud provider	In the field: Global fashion retailer

The Sprinter: AI-ambitious, data-poor

An AI mandate from the top, pilots in several business units, and no owner for the data under any of them, usually on a mid-migration estate. Proxy: engagements with a large or fundamental ambition-to-data gap, 38% of the sample, most with no structural owner for AI readiness. The hidden cost is repetition: each pilot rediscovers the same undocumented definitions at pilot cost, so the same preparation is done several times, in several budgets, and appears in none of them.

First move: name an owner and pick one use case, and pause the parallel pilots long enough to make that use case's foundation work visible, funded, and reusable.

In the field: a UK financial provider arrived with roughly 15 machine learning models, each with its own data preparation, a manual deployment pipeline, and an internal program on hold. The rebuild is the first case study in Part III.[11]

The BI-Locked: a modern estate for the wrong consumer

Modernization is complete: a current platform, curated tables, governed dashboards. The organization reads its maturity as AI readiness, and its roadmap treats the model as a smarter dashboard. Proxy: modern, BI-only platforms, 26% of the sample, where the gap was almost universal and the foundation burden the heaviest. The hidden cost is a second, unplanned modernization: granular history the warehouse aggregated away, unstructured content it never ingested, definitions that lived in analysts' heads and now have to be written down for a model. This is the archetype most surprised by the gap, because it did the work it was told to do.

First move: audit the estate against the AI-ready definition instead of the BI one.

In the field: an e-commerce and subscription platform asked for a better subscription report in Power BI; the engagement delivered a data strategy, an AWS platform, and governance before the dashboard.[12]

The Gated

Often the most technically mature of the four: a modern platform and a serious governance function, built as a control layer for a slower era, so every AI use case queues behind access reviews designed for quarterly reporting changes. Proxy: engagements where the biggest early blocker was governance or access approvals, a third of the sample; about half sat on platforms rated modern and AI-ready. The hidden cost is that caution compounds: each delayed use case confirms AI is risky, justifying another control, while business units find unsanctioned routes to the same outcome.

First move: scope governance per use case, before the pipeline is built, with legal, security, and business owners in the same room, and convert gates to guardrails.

In the field: a global managed cloud provider ran nearly all of its data on-premises in silos, creating inefficient governance. Consolidating more than 70 sources, four warehouses, and a data lake into one governed warehouse removed 17,000 work hours a year and saved over $1M.[13]

The Ready Minority: what they did differently

One engagement in seven in our survey began with the client's AI ambition matched by its data; published research puts the share of enterprises whose data is ready for advanced AI at around 7%.[2] Every one still spent time on foundations, because the work was expected. Most had a timeline that budgeted for data preparation and named data owners from day one; about half had an executive sponsor and documented definitions; one cited reliable pipelines. The group is small, so read this as a pattern to test, not a measurement. The Ready Minority did not buy a different platform. Someone owned the data, the terms were written down, the first use case was narrow enough to judge, and the plan admitted that preparation would take time.

In the field: a global fashion retailer with an existing forecasting model and a working data platform asked for better first-week sales forecasts for new items. The work was model iteration on a sound foundation, and forecast accuracy improved by up to 50%.[14]

Using the archetypes

The archetypes describe where an organization is today; they are not stages it moves through in order. A Sprinter does not have to become BI-Locked before it becomes Ready, and a Gated organization may be one governance redesign away from it. What the four share is that the distance to readiness is organizational before it is technical.

Part III

Closing the gap

The first parts of the Data Index report establish where the gap sits and why it survives. This part closes it, drawing on the same evidence: the conditions our engineers found on the projects that went well, the failures they meet most often, and the anti-patterns our clients describe. The sections follow the order a decision-maker meets them: the strategy document, the foundation it calls for, the governance that lets data flow through it, the sequence in which to build, and two worked examples.

From data strategy to AI-data strategy

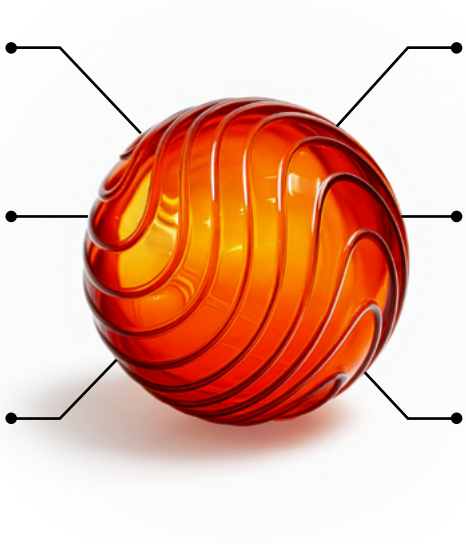
Most data strategies do not need to be thrown out. They need rewriting, because they were written for analysts reading dashboards and now have to serve models and agents reading data directly. Six things change in the document itself.

Six changes to the strategy document

Fund on evidence
First use case plus its foundations; more at gates

Name the ownerscope
One line that assigns accountability for readiness

Put unstructured data in scope
Documents, tickets, and logs are the primary GenAI input



Plan from the use case back
Start from named use cases and work backward to the data

Define readiness per use case
Stated conditions: assets, quality, cadence, owner, definitions

Treat data as products
An owner, a definition, a quality contract, a lifecycle

01 Plan from the use case back

The BI-era roadmap runs the platform first: migrate, consolidate, model, report, then add AI. The AI-data strategy starts with a small number of named AI use cases, asks what data each needs and in what condition, and works backward to the required foundation work. On platform-first programs, the gap survives the migration intact, because nobody asked what the data had to be ready for.

02 Define readiness per use case, and write it down

"AI-ready" becomes a set of stated conditions per use case: which data assets, at what quality thresholds, with what refresh cadence, under whose ownership, with which definitions agreed. The five conditions in section 3 are the template. Analyst guidance on AI-ready data points the same way: align data to the use case and settle governance requirements before the pipeline is built.[6]

03 Treat data as products

A project delivers a pipeline and disbands, whereas a data product has an owner, a definition, a quality contract, consumers, and a lifecycle. The two failures our engineers meet most often, undocumented logic and conflicting sources of truth, are what a product owner exists to prevent.

04 Put unstructured data in scope

BI-era strategies scope the structured estate and leave documents, communications, tickets, and logs out of bounds. For GenAI, they are the primary input. The AI-data strategy names the unstructured sources that matter for its use cases, assigns owners, and budgets for the work our engineers find hardest: deduplication, access, and evaluation.

05 Name the owner

Nearly two-thirds of the engagements our engineers describe began with no structural owner for AI readiness. The strategy fixes this in one line by assigning accountability for each data product's readiness, its definitions, and the currency of the context AI systems reason over. A strategy without a named owner has delegated the work to whichever AI team gets there first.

06 Fund on evidence

Instead of funding a multi-year foundation program on projected value, the AI-data strategy funds the first narrow use case together with its foundation work, sets measurable gates (data assets meeting their stated conditions, a use case in production, a quality incident rate), and releases further investment as gates are passed. On the projects our engineers saw go well, a timeline that budgeted for preparation (28%) and an executive sponsor who accepted foundation work as AI work (26%) were among the conditions in place from day one (Finding 8).

“

Mayank Verma,

VP of Digital Product Management, American Express

We gave identity and verification data named owners, SLAs, and versioned contracts instead of treating it as plumbing. Concrete example: in new-account onboarding, the identity signals feeding our decisioning, document verification, device risk, and KYC attributes are now governed like a product line. One team owns the quality bar, downstream teams consume them through contracts, and a bad data release breaks a build the same way a bad code release would. The most important change was naming an owner who gets paged when data quality degrades.

Dimension	BI-era data strategy	AI-data strategy
Starting point	Platform and migration	Named use cases, worked backward to data
Definition of ready	Platform deployed, dashboards live	Stated conditions per use case: assets, quality, cadence, owner, definitions
Unit of work	Projects that deliver and disband	Data products with owners and quality contracts
Scope	Structured tables	Structured plus the documents, communications, and logs the use cases need
Ownership of readiness	Implicit; falls to the AI team	Named, per data product and per definition
Funding model	Multi-year program on projected value	First use case plus its foundations; further funding at evidence gates

Mitesh Shah,

Product Lead for AI, PayPal

When we scaled AI into production, our models kept degrading and the cause was almost always upstream data, not the models: an event schema changed silently, a pipeline's freshness slipped, or a feature pipeline built for one model got reused by five others. So we started treating key datasets like shipped products. Each has a named owner, an explicit schema contract with versioning, freshness and quality SLAs, and consumers can see the health the way they would a service dashboard. When a schema breaks, the owner is on the hook, not the team that discovers it.

The AI-ready data foundation

What has to exist underneath an AI use case for the data to meet the five conditions, in business language rather than vendor terminology. The estates our engineers walk into usually already have the tooling; what they lack is layers that are present, connected, and owned. Five layers, in the order the field evidence suggests they matter.

Discipline at the source	Enforced validation where records are created was the second most wanted fix among our engineers (30%). It is the one layer that needs business-process change rather than engineering, which is why it is so often skipped.
The semantic layer	The written, maintained agreement on what the business's terms mean, which source is authoritative for each, and how conflicts are resolved, in a form both people and models can consume. It is the fix engineers want most (33%) and find least. No platform ships with it, because it is an organizational agreement expressed as data.
Quality gates in the pipeline	Automated tests on freshness, completeness, distribution, and referential integrity at each stage, with failures that stop data from reaching a model instead of reports describing the damage afterward. The design decision is to make the gates blocking and to test against the semantic layer's definitions as well as the schema.
Active metadata and lineage	With governance encoded in it. Lineage captured automatically from pipelines, definitions linked from the semantic layer, quality status attached to each asset, and policies on permitted use, sensitivity, and retention expressed as rules the platform enforces, so a new use case inherits the policies of the assets it touches. We worked with a Fortune 500 manufacturer to automate lineage documentation through dbt during a Databricks migration, and they absorbed a fivefold increase in data volume with a 10% increase in operating cost.[15]
Unstructured readiness, including evaluation.	A governed corpus with duplicates and superseded versions resolved, sensitive content handled, and access rights carried through into retrieval. The piece our engineers single out as hardest is evaluation (61%): a maintained test set and metrics that say whether retrieval is good enough before an agent acts on it.

The order is roughly the order of payoff. The first three layers fix the failures our engineers meet most often and can be built for one use case within a quarter. Layer four makes that work reusable for the second and third use case, which is the difference between a Sprinter and a Ready Minority organization. Layer five is mandatory for any GenAI use case. Real-time serving paths are built only when a use case demands them (section 15).

Farhat Siddiqui,

Head of Generative AI Enablement & Transformation, Genmab

I would not assign percentages without testing the failure modes separately. An incorrect answer is a symptom. I would ask: Does the source contain a current, authoritative answer? Did retrieval deliver the evidence needed to answer the question? Could the model answer correctly when supplied with that evidence directly? Missing or contradictory source material points to a data problem.

Governance that accelerates instead of blocking

Governance is the dimension where the field evidence is most uncomfortable. The engagements our engineers describe as blocked by governance and access approvals were at least as likely to sit on modern, AI-ready platforms as on anything older. Data leaders know the model has to change; 88% agree that AI demands entirely new approaches to governance and security.[1] Resolving the trust paradox means changing what governance is for.

Guardrails, not gates

A gate stops every use case until a committee reviews it, while a guardrail defines in advance what is permitted for a class of data and a class of use, letting anything inside the boundary proceed. Governance built as guardrails asks the hard questions once, per asset and per use type, and encodes the answers in the platform. The review function then handles exceptions, a workload it can staff, rather than the whole flow, which it cannot.

Gates vs. guardrails, and the one-day session



Every use case waits; months of sequential approvals

Permitted uses defined once per asset and use type; anything inside the boundary proceeds

The governance session One day · all owners in one room: legal, security, data, business

Which data? **For what purpose?** **Under what controls?** **Who signs off?**

Output: asset-level governance for the use case

Scope governance per use case, before the pipeline

The most reliable pattern in our successful engagements is a single working session, before any pipeline is built, with the legal, security, data, and business owners of a use case in one room, answering four questions: which data, for what purpose, under what controls, and who signs off. It takes a day and replaces the months of sequential approvals our engineers report as the biggest early blocker in one engagement in six, and it produces the asset-level governance the AI-ready definition requires as a by-product.

The ownership split that scales

Our Engineering Index found that AI adoption in software delivery scaled only when a central team owned the framework and standards while delivery teams owned enforcement and results. Data governance follows the same rule. A central data office should own the standards: the semantic layer's structure, the policy catalog, the quality-gate patterns, the metadata platform. The domains that produce and consume data should own the content: the definitions of their terms, the quality of their products, the decisions when sources conflict. A central team that tries to own the definitions becomes the bottleneck the Gated archetype describes. Domains that own everything produce the conflicting sources of truth our engineers meet in one engagement in four.

Who owns what

Central data office owns the standards	Domains own the content
Semantic layer structure	Definitions of their terms
Policy catalog	Quality of their products
Quality-gate patterns	Decisions when sources conflict
Metadata platform	

Failure mode: central owns the definitions too → a bottleneck (the Gated archetype)

Failure mode: domains own everything → conflicting sources of truth (1 engagement in 4)



Milos Rusic,

Co-founder and CEO, deepset

Gates are governance theater for agents. A policy document or an approval committee cannot stop a production agent that decides to call an API hundreds of times in a loop. Only controls inside the software can. Start tight and loosen constraints as the system earns trust. Relaxing constraints on a working system is easy. Bolting them onto a fragile one after an incident is far harder.

Someone must own the context

AI systems reason over context: the definitions, documents, and current state they are given. Keeping that context accurate and current is continuous work, and in the BI-era operating model it belongs to no one. Our polling suggests practitioners are willing to assign the role to a dedicated function, though the discipline does not yet have a settled name, and a quarter of organizations have not assigned it at all (Signal 3). The AI-data strategy should assign it explicitly.

Practitioner pulse.

Who keeps your AI's context accurate and current? 30% of practitioners named data and IT teams, 38% a dedicated context-operations function, and 23% said nobody yet.

N-iX audience polling; see Part II.

Anil Yellepeddi,

Senior IT Manager, Applications, Niagara Bottling, LLC

Our governance rule is simple, dollar thresholds drive human approval, and the AI never overrides that. The classifier suggests; the approval hierarchy decides. That was intentional, we wanted the AI to move fast inside a boundary rather than sit behind a review gate that would have killed adoption. Governance that slows everything down teaches the business to route around it. Ours was built to let the AI run where it's safe and stop it where it matters.

Sequencing: the first 90 days vs. the structural work

The most common sequencing failure we see is the eighteen-month data platform program with no AI use case in production at the end of it. The alternative is to sequence the structural work behind a first result, so each quarter produces something in production and something reusable. The Ready Minority in our survey had as much foundation work as everyone else, and a timeline that admitted it.

The first 90 days

One objective: a single AI use case in production on data that meets its stated readiness conditions, with the preparation made visible and reusable. There are six moves, in order.

01 Name the owner	One accountable person for the readiness of the data this use case depends on.
02 Pick one narrow use case	Narrow enough that its data can be enumerated and its success criteria agreed in writing. A third of the successful projects our engineers describe had this from day one.
03 Write the definitions	For the data this use case touches, document what each term means and which source is authoritative. This is the semantic layer, started at the scale of one use case.
04 Put gates on the pipeline	Automated quality checks block anything that doesn't meet those definitions.
05 Hold the governance session	One day, all owners, four questions, and the answers encoded in the platform before the pipeline is built.
06 Ship, and measure	Track the use case in production, the quality incident rate, and the hours spent on foundation work so that the next use case can budget for them honestly.

Venkat Navneeth Burla,

Senior Data Scientist, Vertex Pharmaceuticals

A propensity-based model I worked on looked unusually strong in validation, it had clean lift, good separation across deciles. That's usually a red flag, not a win. When I did a deep RCA, I found a feature that had been derived from a join across two vendor datasets with different timestamp logic. The feature was encoding information that, in the real world, wouldn't exist at prediction time. It only existed because of how the historical data had been joined. We caught it about two weeks before deployment. Fixing it meant rebuilding how validation was handled, moving the temporal split from the model notebook into the pipeline itself so it would catch this class of problem automatically going forward.

The structural work

Behind the first use case, the multi-quarter commitments run in parallel and build on it. The semantic layer expands from one use case's terms to the enterprise's core entities; the metadata platform makes the first use case's definitions, lineage, and quality status queryable; and its policies become the policy catalog. The operating model change, staffing data teams for products instead of backlogs, is the slowest of all and should start in the first quarter even though it will not finish in the fourth. Each archetype's first move (section 8) is the entry point into this sequence.

Archetype	Presenting symptom	First move
The Sprinter	Pilots everywhere, foundations nowhere, no owner	Name the owner; pause parallel pilots long enough to make one use case's foundation work visible and reusable
The BI-Locked	Modern platform, universal gap, heaviest foundation burden	Audit against the AI-ready definition; scope the unstructured estate and the definitions analysts carry in their heads
The Gated	Sound estate, nothing reaches a model	Governance session per use case; convert gates to guardrails; adopt the central-standards, domain-content split
The Ready Minority	Data has stopped being the constraint	Institutionalize: metadata platform, policy catalog, context ownership; move from one use case to a portfolio



Case studies: closing the gap in practice

Two engagements, one on structured data and one on unstructured, illustrate the points highlighted in the Data Index. [7] [8]

Case 1: From fifteen orphaned models to a governed decision pipeline

Starting point. A UK provider of prepaid cards and current accounts, with over a million customers and billions of pounds in processed payments, needed to process payments instantly while applying anti-money-laundering, sanctions, no clear way to call the models, deployment was entirely manual, and an internal effort to fix this had been put on hold for lack of a clear target design. A Sprinter, in the terms of this report: ambition well ahead of the data and no owner of the decision pipeline. The AI already existed: nearly 15 machine learning models sat across multiple systems, each managed separately with its own data preparation. Models recalculated every five minutes, so decisions waited. Other system components had no clear way to call ahead of the data and no owner of the decision pipeline.

What changed. The engagement began with a discovery phase that produced the roadmap and budget the internal effort had lacked. The foundation work followed the order: a feature store with API endpoints, so every model drew from one governed feature store instead of its own preparation path; change data capture from the source systems into a streaming layer for real-time feature calculation; a model registry and a standard runtime so deployment stopped being manual; and a clean split between batch and real-time data flows. Datasets were anonymized for privacy, and the whole stack was built cloud-agnostic.

Outcomes. Transaction processing latency fell from up to five minutes to 250 milliseconds at loads of up to a million requests a month. The central SQL data server was retired, reducing operating costs. The client reported a 20% increase in customers and a more than 30-point rise in its Net Promoter Score over the period.

Metric	Before	After
Transaction decision latency	Up to 5 minutes	250 milliseconds
Feature management	Separate per model, ~15 models	One feature store with API endpoints; per-model ETL retained
Deployment	Entirely manual	Registry and standard runtime, cloud-agnostic
Customer growth over the period	—	+20%; NPS up more than 30 points

Case 2: Documenting the estate before retrieving from it

Starting point. A UK enterprise software vendor with more than 600 clients wanted to raise developer productivity for one of its customers by improving documentation coverage of enterprise-architecture components and processes, increasing visibility into which applications were used where, and enabling faster retrieval from an extensive knowledge base. Documentation of back-end services and processes was patchy, application-to-process mappings weren't maintained, and knowledge base search was manual and slow. Retrieval-augmented generation over that corpus would have returned the gaps.

What changed. The team built two things together on Azure AI Foundry. Applications that generated descriptions for undocumented services and processes, and automatically detected application-to-process usage, turned the estate's tacit knowledge into a maintained corpus. Alongside them, a RAG-based assistant over Azure AI Search, with on-demand chart generation from the underlying data, gave developers a way to retrieve from that corpus. The generated documentation made the assistant's answers trustworthy: the retrieval and semantic layers for the unstructured estate were built as one piece of work.

Outcomes. Knowledge base search became roughly 120 times faster. Generating a data-driven chart went from hours of development effort to seconds. Users without a technical background could find and work with the data directly.

Metric	Before	After
Knowledge base search	Manual, slow	~120x faster
Data-driven chart generation	Hours of development	Seconds, on demand
Documentation of back-end services and processes	Partial	Generated automatically
Access for non-technical users	Through engineers	Direct

The two cases differ in every technical respect and share a principle. In both, the work the client asked for depended on the underlying data having an owner, a definition, and a way to be trusted.

Closing

Self-diagnostic: the AI-ready data scorecard

The scorecard turns the five dimensions into a readiness profile. Complete it for one named AI use case, with the people closest to the data in the room; Part I showed that a scorecard completed in the boardroom will overstate readiness by roughly the margin this report is about. There are four statements per dimension; count the ones that are true today.

Quality and consistency

- For the data this use case depends on, we can name the authoritative source for each core entity, and nobody in the business disputes it.
- Business logic in our transformations is documented well enough that a new engineer could explain what each pipeline does without reading the code.
- Quality checks run automatically in the pipeline, and a failed check stops data reaching the model.
- We know our data quality incident rate for this use case, and it is falling.

Access and architecture

- The data this use case needs can be reached and joined without a multi-week integration project.
- The documents, communications, and logs relevant to this use case are in scope of the platform.
- Granular records and event history are available where the use case needs them.
- Where the use case needs current data, a path exists to deliver it at the cadence of the decision.

Governance and trust

- For each data asset this use case touches, the use case defines and records permitted uses, sensitivity, and the owner.
- The platform enforces policies, so new use cases inherit them; review is an exception process.
- Governance requirements were settled before the pipeline was built, in one session with legal, security, data, and business owners present.
- Standards are owned centrally; definitions and quality are owned by the domains that produce the data.

Metadata and context

- Lineage for this use case's data is captured automatically and is current.
- A maintained semantic layer records what our core business terms mean, in a form both people and AI systems can use.
- For GenAI use cases, we have an evaluation set and metrics that tell us whether retrieval is good enough before an output is acted on.
- Someone is accountable for keeping the context our AI systems reason over accurate and current.

Talent and operating model

- A named person owns making this use case's data AI-ready, and it is not the AI project team by default.
- The plan and budget for this use case include the foundation work explicitly, with hours or cost attached.
- An executive sponsor has accepted that foundation work is part of the AI investment.
- Our data team is staffed and measured for data products and quality operations, not only for the reporting backlog.

Reading your score

Statements true	Profile	What it usually means
0 to 8	Not ready	The gap is structural. Start with the first-quarter moves: owner, one use case, definitions, gates, governance session. Expect the first use case to carry most of its own foundation work. If your profile shows a structural gap, talk to our data team.
9 to 15	Partially ready	Foundations exist unevenly. Find the weakest dimension, usually metadata or operating model, and build the reusable layers so the second use case inherits from the first.
16 to 20	AI-ready for this use case	Data is no longer the constraint for this use case. Institutionalize: metadata platform, policy catalog, context ownership. Repeat the scorecard for the next use case; readiness does not transfer automatically.

A total conceals the profile, and the profile determines the first move: an organization scoring 14 with zeros in governance is Gated, and one scoring 14 with zeros in metadata needs a different intervention entirely. Enterprises rarely fail on one dimension, and the dimensions differ in remediation horizon: quality and metadata respond to focused engineering within a quarter, while architecture and operating model take several. That asymmetry is why a profile is more useful than a score.

2027 outlook: from AI-ready to agent-ready

Everything in this report assumes AI that produces an output a human reviews. Agents retrieve data across systems, chain decisions, and act, and the market is committing to them faster than it is preparing for them: nearly half (48%) of the practitioners we polled expect multi-agent collaboration to deliver the most enterprise value over the next three years (Signal 4).

The readiness gap for agents is the gap described in this report, raised on three axes. Agents need semantic context, a machine-readable account of what the business's terms mean and how its entities relate, which is the layer most estates have not built. They need currency, data refreshed at the speed of action, which most pipelines weren't designed for: event streams and serving layers that deliver granular records in seconds, built for the use case that needs them, not before. And they need governance that runs at machine speed, enforced on every retrieval and every action, because no human is in the loop to catch exceptions. An estate that is ready for a model a human reviews is not automatically ready for an agent, and the difference is exactly the layers our engineers report as least built.

Practitioner pulse.

48% of practitioners expect multi-agent collaboration to deliver the most enterprise value within three years, and 47% see agents' most practical analytics value in prediction.

N-iX audience polling; see Part II.

Analysts project that a large share of agentic AI projects will be canceled by the end of 2027, and a readiness benchmark published this year found far more organizations running agentic AI in production than are prepared for it, with data quality and lineage the most cited barrier.[18] [19] Those findings describe the same organizations: deploying agents on estates that were not ready for the models before them.

From AI-ready to agent-ready: the same gap on three axes



Semantic context
Machine-readable meaning of terms and entities



Currency
Data refreshed at the speed of the action



Governance at machine speed
Enforced on every retrieval and action

48%
expect multi-agent collaboration to deliver the most value in three years
N-iX audience polls

48%+
of agentic AI projects will be canceled by the end of 2027
Gartner

15% vs 41%
fully prepared for agentic AI vs. already running it in production
Fivetran

What changes in 2027 planning

The semantic layer moves from useful to prerequisite; an agent without shared business meaning is a retrieval system with permission to act. Context becomes an operated asset with a named owner, whatever the function is eventually called. Evaluation becomes non-negotiable, because the problem our GenAI engineers report as hardest, knowing whether retrieval is good enough, is the problem that decides whether an agent can be trusted. And governance has to be encoded in the platform, because at agent cadence a review queue becomes an incentive to route around it.

This report recommends the same decisions that make an estate agent-ready: an owner, a use case, written definitions, blocking quality gates, and governance settled in advance. Organizations that make them in 2027 for the AI they run today will find the agents they deploy in 2028 have somewhere to stand.

Tural Hagverdiyev,

Software Engineer, Meta

Before an autonomous agent is allowed to act on enterprise data, there must be an explicit, deterministic authorization boundary checked at the point of action. What the agent is allowed to do must be enforced by the environment.

An agent that can read enterprise data should need a verifiable grant to act on it: scoped to a specific task, checked every time, and recorded in an audit trail. Until that boundary is mechanical and testable, every agent deployment is one confused or compromised plan away from a data incident.



Wrap-up

The gap this report measures sits in decisions, not in platforms. Forty percent of the estates our engineers walk into are modern, and two-thirds of those still cannot support the AI their owners want. The gap is in definitions no one wrote down, ownership no one assigned, governance designed for a slower consumer, and unstructured knowledge no one brought into scope. Each of those is a decision, and every organization in our sample could have made it before the first engineer arrived.

The organizations in the ready minority did not outspend the others; they named an owner, agreed what their terms meant, chose a first use case narrow enough to judge, and wrote a plan that admitted the foundation work would take time. When their AI initiatives stalled, they were the group most likely to say the data was fine. The bar is rising: agents will consume context, currency, and governance most estates have not built, at a speed that exposes every gap described here. The scorecard in section 14 is where to start.

About N-iX Data & Analytics

N-iX is a strategic partner in Data, AI & Analytics for enterprises, helping to move from AI experimentation to measurable production outcomes. We support companies in reinventing their data strategy, from data modernization to AI adoption at scale. As a Snowflake Elite Partner and a Premier Tier AWS Partner, N-iX combines deep platform expertise with a broad ecosystem of AI technologies, including partnerships with Anthropic (Claude), OpenAI, and xAI (Grok).

2,400+
Tech experts

90+
Fortune 500 clients

300+
Data projects delivered

Take the next step

Complete the scorecard for one named AI use case, with the people closest to your data in the room. If the profile shows a structural gap, N-iX Data & Analytics offers a readiness assessment that maps your estate against the five conditions and the five layers and produces a sequenced first-quarter plan.

Lead Author:



Rostyslav Fedynyshyn

Head of Data & Analytics Practice, N-iX

Rostyslav drives data modernization, analytics, and business intelligence initiatives for enterprise clients. His focus areas include data platform architecture, cloud data solutions, and advanced analytics across industries. He brings over 13 years of experience in data engineering, analytics, AI, and technology consulting.

Book a consultation with Rostyslav

Contributing Authors:



Danylo Diachenko

Senior Data Architect



Maryana Sydoriv

Market Intelligence Senior Manager

contact@n-ix.com
n-ix.com
+44 20 3740 7669
+1 727 341 5669

Appendix A.

Audience polls: full results

Fourteen closed single-select polls run across N-iX professional channels between September 2025 and August 2026, 3,104 responses in total. Percentages are of respondents to each poll.[9]

Question	Responses
Roughly what percentage of the total AI lifecycle is typically spent on data preparation?	<20%: 12% · 30–50%: 33% · 60–80%: 46% · >90%: 6%
Which part of the AI lifecycle actually eats up 60–80% of the time?	Organizing data: 50% · Feature engineering: 20% · Model tuning: 21% · Deployment: 8%
When is bias most effectively and cheaply corrected in an AI project?	During data prep: 48% · Model tuning phase: 27% · Model selection: 21% · Post-deployment: 2%
Which AI strategy will deliver the most enterprise value over the next 3 years?	Multi-agent collaboration: 48% · Domain-specific models: 25% · Physical AI and robotics: 14% · General multi-purpose LLMs: 11%
What's the biggest challenge to GenAI adoption?	Data literacy: 48% · Cloud scalability: 22% · Lack of funding: 14% · Inefficient data pipelines: 14%
What are the biggest challenges your organization faces when implementing new technologies?	Finding skilled professionals: 62% · Managing large amounts of data: 20% · Dealing with inaccuracies: 16%
How has AI coding impacted your engineering organization so far?	Faster dev, same release: 52% · Faster releases: 34% · Too early to measure: 13%
When scaling GenAI, what's the #1 data gap?	Limited data quality: 47% · Data silos and fragmentation: 26% · Unstructured data: 21% · Legacy data stacks: 4%
AI agents for enterprise data analytics: where do you see the most practical value?	Predictive analytics: 47% · Real-time KPI monitoring: 21% · Self-service analytics: 17% · Other: 13%
What is the biggest barrier to widespread adoption of AI agents in enterprise settings?	Data security risks: 85% · Complexity of implementation: 4% · Talent and skill shortage: 4% · Other: 4%
What is the main cause of GenAI hallucinations?	Model's limited knowledge: 45% · Lack of real-time data: 20% · Weak prompt design: 20% · Poorly managed data: 15%
What is the biggest challenge in adopting open-source AI?	Cybersecurity and data privacy: 64% · Expertise and time-to-value: 28% · Regulatory and IP issues: 7%
What's the biggest barrier to executing your AI strategy? (Reddit)	AI literacy and skills · Governance and structure · Data and tech readiness · Funding and investment (ranked; small sample)
Who is responsible for keeping your AI's context current?	Dedicated ContextOps/FinOps: 38% · Data and IT teams: 30% · Nobody yet: 23% · Prompt engineers: 7%

Appendix B.

N-iX Data and AI Engineering Survey: headline results

Anonymous questionnaire completed by 258 N-iX data and AI practitioners in September 2026 about their most recent client engagement. Percentages are of all respondents unless stated; multi-select questions report the share of respondents selecting each option.[10]

Question	Responses
Client's data platform on arrival	Modern, AI-ready: 40% · Mid-migration hybrid: 30% · Modern, designed for BI only: 26% · Mostly legacy / on-prem: 2%
Distance between client AI ambition and what the data could support	Modest gap (weeks): 42% · Large gap (months): 33% · Ambition matched the data: 14% · Fundamental gap: 5% · Ambition not defined specifically enough: 5%
Single biggest blocker in the first 90 days	Data quality and consistency: 26% · Governance and unclear ownership: 16% · Access and security approvals: 16% · Unclear business requirements: 16% · No significant blockers: 16% · Silos and legacy architecture: 7% · Missing metadata and lineage: 2%
Share of time spent fixing foundations rather than delivering new value	Under 25%: 47% · 25–50%: 30% · 50–75%: 16% · Over 75%: 7%
Share of critical data problems the client had documented beforehand	Almost none (0–10%): 26% · Some (11–33%): 49% · Many (34–66%): 19% · Most (67%+): 2%
Share of total effort on data preparation	Less than 30%: 21% · 30–50%: 40% · 51–70%: 21% · More than 70%: 9% · Hard to separate: 7%
Data quality failure hits most often, career-wide	Undocumented business logic in transformations: 54% · Conflicting sources of truth: 28% · Incomplete or missing records: 14% · Silent schema drift: 5%
One thing to fix before data reaches the pipeline	Consistent, documented business definitions: 33% · Enforced data entry validation at source: 30% · Clear data ownership with a named person: 16% · Complete event history: 12% · Other: 9%
Worked on a GenAI/RAG/LLM project requiring unstructured data preparation	Yes, most recent engagement: 30% · Yes, earlier engagement: 12% · No, not yet: 58%
Hardest part of making unstructured data usable (multi-select; respondents with GenAI experience)	Evaluation: 61% · Deduplication and conflicting versions: 44% · Getting access to documents: 33% · Embedding and retrieval quality: 33% · Chunking strategy: 22% · Detecting sensitive information: 22%
Who owns making data AI-ready at the typical client	The AI project team, ad hoc: 35% · Nobody clearly owns it: 28% · Central data/platform team: 28% · A named data or product owner: 9%
Most common data-related reason an AI initiative stalls	Data exists but quality too poor to trust: 44% · Data was fine, failure elsewhere: 28% · Compliance, privacy, or sovereignty constraints: 16% · Data siloed, no owner of source of truth: 9%
What's the biggest barrier to executing your AI strategy? (Reddit)	AI literacy and skills · Governance and structure · Data and tech readiness · Funding and investment (ranked; small sample)
Who is responsible for keeping your AI's context current?	Dedicated ContextOps/FinOps: 38% · Data and IT teams: 30% · Nobody yet: 23% · Prompt engineers: 7%
What was in place from day one on projects that went well (multi-select)	Named data owners who could make decisions: 58% · Data access and documented definitions before kickoff: 53% · Narrow, well-defined first use case: 35% · Reliable existing pipelines and quality checks: 28% · Timeline that budgeted for data preparation: 28% · Executive sponsor accepting foundation work as AI work: 26%

References and methodology notes

A full list of the external research consulted for this report is given here, together with the methodology notes for N-iX's own evidence. Superscript numbers in the body refer to this list.

Reference set consulted: Salesforce, State of Data and Analytics, November 2025; Cisco, AI Readiness Index, 2025; Informatica, CDO Insights 2026, January 2026; Cloudera, The Data Readiness Index, April 2026; Harvard Business Review Analytic Services and Hyland, Bridging the Readiness Gap to the Agentic Enterprise, April 2026; Fivetran, The 2026 Agentic AI Readiness Index, May 2026; Drexel University LeBow College of Business and Precisely, 2026 State of Data Integrity and AI Readiness, January 2026; Accenture, AI-Ready Data, May 2026; Gartner, Lack of AI-Ready Data Puts AI Projects at Risk, February 2025; Gartner, Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, June 2025.

1 Salesforce, State of Data and Analytics, November 2025. Double-anonymous surveys of 3,800 analytics and IT decision-makers and 3,852 line-of-business leaders across 18 countries. https://www.salesforce.com/news/stories/data-analytics-trends-2026/

2 Accenture, AI-Ready Data: New Rules of Data for the Advanced AI Era, May 2026. Survey of executives at 2,000 companies across 15 countries and 9 industries. https://www.accenture.com/us-en/insights/ai-data/ai-ready-data

3 N-iX pipeline analysis: 185+ enterprise data and analytics opportunities recorded in N-iX's CRM over the review period, classified from the client's own request description into non-exclusive demand patterns (mean 1.4 patterns per opportunity) and into solution sub-clusters. Shares and conversion figures are held internally.

4 N-iX Data and AI Engineering Survey, September 2026. Anonymous questionnaire completed by 258 N-iX practitioners: 47% data engineers, 23% ML or AI engineers, the remainder architects, DevOps, delivery, and product roles. 60% have more than six years of experience in technology. Most recent engagements: healthcare and life sciences 35%, retail and e-commerce 19%, logistics 14%, financial services 9%, other 23%. Client size: 58% above 5,000 employees, 19% below 500. Multi-select questions report the share of respondents selecting each option. Sub-group figures are based on 7 to 21 respondents and are reported as indicative.

5 N-iX audience polling, September 2025 to August 2026: 47 closed single-select polls across N-iX professional channels (LinkedIn and Reddit), 1,174 total responses. The 14 polls addressing data readiness directly are analyzed in this report: 3104 responses, 13 to 86 per question, median 23. Respondents are self-selected members of the N-iX professional audience, predominantly data engineering, analytics, AI/ML, and technology leadership roles.

6 Gartner, Lack of AI-Ready Data Puts AI Projects at Risk, February 2025. Survey of 1,203 data management leaders, July 2024. https://www.gartner.com/en/newsroom/press-releases/2025-02-26-lack-of-ai-ready-data-puts-ai-projects-at-risk

7 Harvard Business Review Analytic Services, sponsored by Hyland, Bridging the Readiness Gap to the Agentic Enterprise, April 2026. Survey of 325 HBR audience members, fielded December 2025. https://www.hyland.com/en/company/newsroom/new-harvard-business-review-analytic-services-research-exposes-ai-gap

8 Cloudera, Nearly 80% of Enterprises Say AI Is Held Back by Data Access Challenges, April 2026. Survey of 1,270 IT leaders at companies with more than 1,000 employees. https://www.cloudera.com/about/news-and-blogs/press-releases/2026-04-14-nearly-80-percent-of-enterprises-say-ai-is-held-back-by-data-access-challenges-cloudera-report-finds.html

9 Informatica, CDO Insights 2026, January 2026. Survey of 600 data leaders. https://www.informatica.com/about-us/news/news-releases/2026/01/20260127-new-global-cdo-report-reveals-data-governance-and-ai-literacy-as-key-accelerators-in-ai-adoption.html

10 Drexel University LeBow College of Business and Precisely, 2026 State of Data Integrity and AI Readiness, January 2026. Survey of 505 data and analytics leaders. https://www.lebow.drexel.edu/sites/default/files/2026-01/lebow-precisely-state-data-integrity-ai-readiness-2026.pdf

11 N-iX, Cloud-agnostic ML solution fuels 20% customer growth for a UK financial provider. https://www.n-ix.com/case-study/customer-growth-finance-machine-learning/

12 N-iX, Driving growth in e-commerce with a comprehensive data analytics solution. https://www.n-ix.com/case-study/e-commerce-comprehensive-data-analytics-solution/

13 N-iX, Managed cloud provider saves $1M+ and cuts 17,000 work hours with automation and cloud migration. https://www.n-ix.com/case-study/automation-cloud-migration-cost-optimization-for-global-tech-company/

14 N-iX, How a global fashion retailer improved sales forecasting by 50%. https://www.n-ix.com/case-study/improving-product-allocation-retail-sales-forecasting/

15 N-iX, Reducing data processing costs for a Fortune 500 manufacturing leader scaling to 5x data volume. https://www.n-ix.com/case-study/databricks-migration-fortune-500-technology-company/

16 N-iX, Engineering Index 2026: Pragmatic AI Software Engineering. A Framework for CTOs and Engineering Leaders, 2026.

17 N-iX, Enterprise software leader makes knowledge base search 120x faster with AI. https://www.n-ix.com/case-study/ai-powered-knowledge-search-enterprise-software-leader/

18 Gartner, Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, June 2025. Poll of 3,412 webinar attendees, January 2025. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

19 Fivetran, The 2026 Agentic AI Readiness Index, May 2026. Survey of 400 data professionals. https://www.fivetran.com/resources/reports/the-2026-agentic-ai-readiness-index

20 N-iX Data and AI Engineering Survey (see reference 4). Archetype proxies: Sprinter, engagements reporting a large or fundamental ambition-to-data gap; BI-Locked, platforms rated modern but designed for BI only; Gated, engagements whose biggest first-90-day blocker was governance and ownership or access and security approvals; Ready Minority, engagements where ambition matched the data. Groups overlap; shares are indicative.

N-iX

contact@n-ix.com
n-ix.com

+44 20 3740 7669
+1 727 341 5669