MERIT

Closing the AI ROI Gap - How Data Engineering Separates High-Performing Enterprises From the Rest

Most AI initiatives fail not because the model is wrong - but because the data engineering beneath it was never built to scale.

Authored by Tharun Mathew

Head of Data & AI Solutions, Merit Data and Technology

EXECUTIVE SUMMARY

SITUATION

Every business that runs on data faces the same challenge. The data that powers competitive advantage cannot be bought off a shelf. It must be harvested - with precision, at scale, and at speed. In 2026, most organisations are using AI to do that harvesting. Most of them are getting it wrong.

COMPLICATION

And yet, only 25% of organisations have moved 40% or more of their AI pilots into production.[2] The bottleneck isn't the model. It's not the algorithm, the vendor, or even the use case. In the vast majority of cases, it's the data engineering infrastructure underneath - fragmented pipelines, unstructured silos, absent governance, and architectures that were built for reporting, not for real-time intelligence.

RESOLUTION

The organisations whose AI initiatives reach production - and actually deliver commercial returns - share one thing in common: they've moved from pipeline thinking to platform thinking. They treat data engineering not as back-office plumbing, but as the core infrastructure that determines whether AI creates value or stays permanently stuck in pilot mode.

KEY FINDINGS

1. **The pilot-to-production gap is a data engineering failure.** 74% of organisations struggle to scale AI value because of data governance and accessibility problems - not model limitations. The problem is structural, not technical.
2. **Pipelines are necessary but no longer sufficient.** Batch pipelines built for BI and reporting can't support the real-time, governed, context-rich data flows that production AI and agentic systems need.
3. **Platform thinking changes the economics.** Organisations that shift to unified data engineering platforms achieve 3.7x average ROI on data and AI investment. Engineering teams using DataOps practices are ten times more productive than those that don't.
4. **Governance Isn't Compliance. It's Risk Management.** In a world shaped by GDPR, the EU AI Act, and growing litigation risk around automated decisions, governance is no longer a compliance checkbox - it's your first line of defence against untrustworthy AI outputs. 83% of enterprises have already shifted their data management priorities in response to AI. The organisations that embed governance into their pipelines from the outset aren't just building cleaner infrastructure - they're protecting themselves from the decisions their AI will make tomorrow.
5. **The data engineering platform has five interdependent pillars.** Strategy & Architecture, Cloud & Data Platforms, Governance & Compliance, AI & Data Validation, and DevOps for Data & AI. All five need to work together. None is optional.

KEY STATISTICS

25%
of organisations have moved 40%+ of AI pilots into production
(Deloitte, 2026)

74%
struggle to scale AI value due to data governance and access failures
(BCG, 2025)

3.7x
average ROI on data and AI investment for platform-grade infrastructure
(IDC / Microsoft, 2025)

CONTENTS

Table of Contents

Section 1

The AI Pilot Trap: A Data Engineering Problem

Enterprise AI has a scaling problem - and it's not the problem most boards think it is. Enterprise AI has a scaling problem - and it's not the problem most boards think it is. The usual explanations for AI failure - models that hallucinate, use cases that are poorly scoped, teams without the right skills - are real, but they're only part of the story. The most consistent, most costly, and most underappreciated reason AI initiatives stall between pilot and production is more fundamental: the data engineering infrastructure was never designed to support production-grade AI.

This is what pilot purgatory actually costs. Every quarter a promising AI initiative sits in extended pilot, the bill compounds - cloud compute and storage spend on AWS and Azure that delivers no business output, senior data engineers and ML specialists burning time on pipeline firefighting instead of building capability, and vendor licensing fees accumulating on tools that never reach scale. These aren't sunk costs in the accounting sense. They are live, recurring losses that show up quietly across infrastructure budgets and headcount allocation while faster, more agile competitors are already in production.

The IBM Institute for Business Value's 2025 CEO Study found that only 16% of AI initiatives have successfully scaled across the enterprise. BCG research found that 74% of organisations struggle to scale AI value specifically because of data governance and accessibility issues - not model limitations. More than half of surveyed executives reported that difficulties integrating AI infrastructure with legacy systems directly derailed their target outcomes.

These aren't AI problems. They are data engineering problems wearing an AI mask - and switching models or changing vendors won't fix them. What they will do is add another line to the same invoice.

"More than half of surveyed executives said that difficulties integrating AI infrastructure with legacy systems directly derailed their target outcomes."

- IBM Institute for Business Value, 2025

The Anatomy of the Gap

Fragmented pipelines. Most enterprise data environments were built incrementally - pipeline by pipeline, use case by use case - with no unified architecture underneath. The result is a patchwork of batch jobs, point-to-point integrations, and undocumented data flows that can't consistently deliver the governed, real-time data access that AI systems need.

Unstructured data locked in silos. IBM's 2026 data trends analysis estimates that enterprises typically have up to 90% of their data locked in unstructured formats - PDFs, contracts, emails, logs, images - that legacy pipelines were never built to handle.[5] AI models trying to operate across this data have nothing solid to stand on.

Governance gaps that compound downstream. 83% of enterprises have already shifted their data management priorities because of AI demands.[6] But shifting priorities isn't the same as solving the problem. Most organisations still lack lineage tracking, data contracts, and automated quality checks at the pipeline level - the basics that make AI outputs trustworthy and defensible in production.

Batch architectures in a real-time world. Pipelines built for overnight reporting runs can't power pricing engines, fraud detection, or customer intelligence platforms that need data refreshed in seconds or milliseconds. 31% of organisations already report revenue loss directly attributable to data lag or pipeline downtime.[7]

The Compounding Cost of Pilot Purgatory

Every quarter an AI initiative stays in pilot mode, the cost compounds. Engineering time gets consumed maintaining proof-of-concept infrastructure instead of building production systems. Competitive windows close. The organisation's AI capability falls further behind peers who have already crossed the line. The investment is real. The returns are not - and they won't be until the data engineering foundation underneath is rebuilt to support what the business actually needs.

Section 2

What 'AI-Ready' Actually Demands from Infrastructure

The phrase 'AI-ready data' gets thrown around a lot. What it actually demands of data engineering infrastructure is much less well understood. AI-ready isn't a data quality standard. It's an architectural condition. A dataset can be accurate, complete, and reasonably well-governed - and still fail to be AI-ready if it can't be delivered when the system needs it, in the format it requires, at the volume it operates at. Getting there requires rethinking data engineering from the architecture up. Cleaning data at the tail end of existing pipelines won't cut it.

What AI Systems Actually Need from Data Infrastructure

Consistency at scale. Machine learning models are sensitive to inconsistencies in training data, and production AI systems are equally sensitive at the point of decision. Both require pipelines that apply the same transformation logic, quality rules, and schema enforcement across millions of records - with no variance introduced by manual fixes, parallel pipelines, or undocumented exceptions. One inconsistency at scale becomes thousands of bad decisions.

The ability to act while the moment still matters. Agentic AI systems - the frontier of enterprise AI in 2026 - make autonomous decisions in real time. They need data that is fresh enough to be useful now: while the customer is still on the app, while the transaction is still in progress, while the decision still has commercial value. Organisations that get this right see measurable gains in conversion, retention, and fraud prevention. Those that don't are making AI-powered decisions on yesterday's reality.

Semantic context and metadata richness. AI agents don't just need data - they need to understand what it means. Where it came from, how reliable it is, how it relates to other datasets, and what governance rules apply. That requires infrastructure built around lineage, data contracts, and governed catalogues that give AI systems the context to act reliably -not just quickly.

Automated governance at enterprise scale. AI systems working across organisational boundaries need governance frameworks that enforce access controls, privacy rules, and regulatory requirements automatically - continuously, and without human intervention at each data call. Manual governance doesn't scale. At enterprise AI volumes, it simply breaks.

KEY STATISTICS

53%
of executives say AI integration failures with legacy systems derailed target outcomes (IBM IBV, 2025)

83%
of enterprises have shifted data management priorities in response to AI demands (IDC, 2025)

42%
of organisations cannot properly customise AI models due to poor-quality data (IBM, 2025)

Why Most Existing Infrastructure Falls Short

The gap between what AI needs and what most enterprise data infrastructure provides isn't primarily a technology gap. The tools exist. Modern cloud platforms, streaming frameworks, orchestration layers, and governance engines are all mature and commercially proven. The gap is architectural. Most enterprise data environments were designed for a world where the end consumer of data was a human analyst building a report.

That architecture - batch processing, tabular structures, manual quality reviews, siloed storage - is misaligned with the needs of AI systems that consume data autonomously, at scale, in real time, across unstructured as well as structured sources. IBM IBV research found that only 29% of technology leaders strongly agree their enterprise data meets the quality, accessibility, and security standards needed to scale generative AI.[8] That 71% shortfall isn't mainly a data quality problem. It's an infrastructure architecture problem - and you can't clean your way out of it. The engineering platform underneath needs to change.

Section 3

The Platform Imperative: Beyond Pipeline Thinking

A pipeline moves data from A to B. A platform makes data available to every system that needs it - in the right form, at the right time, with the right governance - reliably and at scale. That distinction matters more than it might sound. Organisations that treat data engineering as a series of pipelines to build and maintain end up in perpetual technical debt. Every new AI initiative needs new pipelines, new integrations, new quality checks- built from scratch each time. The work never compounds. The capability never scales. And the engineering team's bandwidth for new AI use cases shrinks as maintenance overhead grows.

This is the hiring trap most organisations don't see coming. As pipeline complexity grows, the instinct is to hire more engineers to manage it. But data engineering talent is scarce, expensive, and increasingly competed for by every enterprise running an AI programme. Organisations that continue down the pipeline path will find themselves in a cycle where headcount grows, costs rise, and delivery speed stays flat - because more engineers managing more pipelines is not the same as more capability.

You can see this pattern in the complexity of most modern data stacks. Organisations now manage five to seven or more specialised data tools on average, with 70% of data leaders reporting that stack complexity is a significant operational challenge. Each tool was chosen to solve a specific problem. Each one added integration overhead. Each integration added fragility. For many organisations, the data stack has quietly shifted from being a solution into being the problem.

"The coordination layer is now the scarce resource - not the code. The best data engineers of 2026 think like site reliability engineers, building platforms that let teams ship data safely, with guardrails that prevent the most common failure modes."

- Kestra, 2026 Data Engineering Trends

What Platform Thinking Changes

From bespoke to reusable. A platform defines common patterns, shared infrastructure, and standardised interfaces that every new data initiative builds on. Instead of rebuilding pipeline logic for each use case, teams configure and extend what's already there. Development time drops. Quality goes up. Governance is inherited by every new workload rather than reinvented each time. Critically, each new AI use case no longer requires a proportional engineering effort to support it.

From reactive to proactive. Pipeline environments are fundamentally reactive - data moves when a scheduled job runs, quality gets checked after the fact, failures surface when dashboards break or models degrade. Platform environments embed observability, automated quality checks, and alerting into the infrastructure itself. Problems get caught at the source, not downstream when the damage is already done - and not by an engineer who could have been building something new.

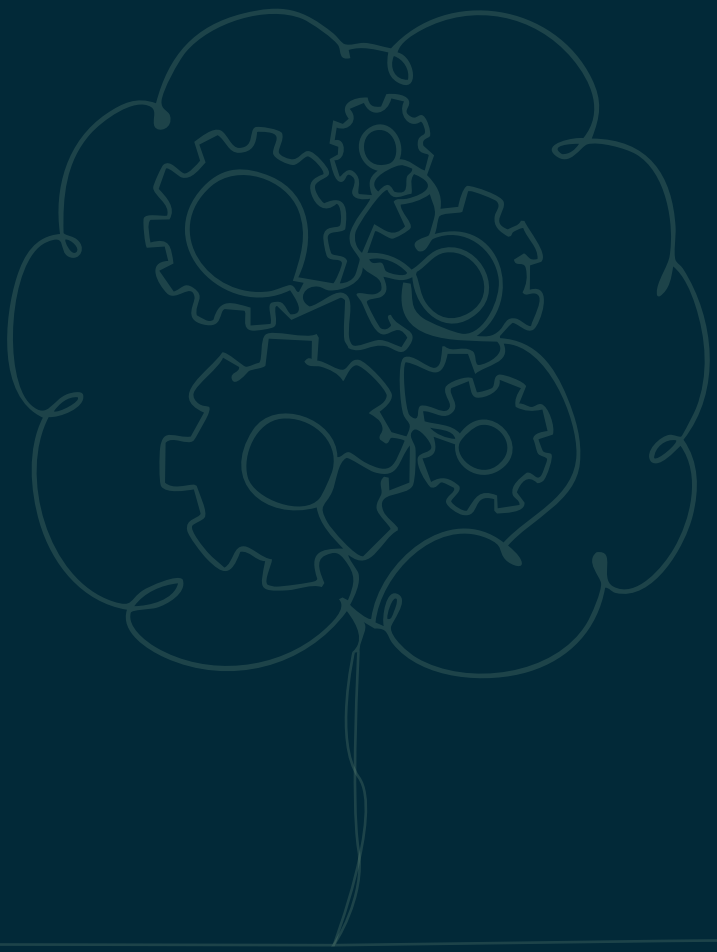
From cost centre to strategic asset. When data engineering is a platform rather than a collection of pipelines, its value compounds with every use case built on it. The governance, quality, and reliability already built into the platform are inherited by everything that runs on it. Over time, the platform becomes a genuine source of competitive advantage - not just the plumbing behind it.

From engineering bottleneck to self-service capability. Platform-grade data engineering includes semantic layers, governed catalogues, and data contracts that let data scientists, analysts, and AI teams find and use trusted data without filing engineering tickets for every new request. Data engineering teams shift from fulfilling requests to stewarding a platform. The practical result: organisations can scale their AI output without scaling their engineering headcount in lockstep. In a market where a senior data engineer costs upwards of £80,000-£120,000 per year and takes months to hire, that is a material financial advantage.

The Economics of Platform vs. Pipeline

IDC research shows organisations achieve 3.7x average ROI on data and AI investments when infrastructure is properly designed - with top performers reaching 10.3x. DataOps-guided engineering teams are projected to be ten times more productive than those without platform-level practices. Modern cloud-native adopters achieve 50% lower total cost of ownership versus fragmented pipeline environments.

The implication for leadership is straightforward: the platform approach does not just reduce technical debt. It reduces the headcount required to deliver AI at scale - and it widens the gap between those who made the transition early and those still managing pipelines by hand. That gap gets more expensive to close every year.



Section 4

Five Pillars of a Modern Data Engineering Platform

A data engineering platform isn't a single technology. It's five interdependent capabilities designed to work as a coherent whole. The pillars aren't sequential build phases. They're simultaneously active architectural concerns that constrain and reinforce each other. Invest in cloud platforms without embedding governance and you can't trust what those platforms deliver. Build governance frameworks without DataOps discipline and you can't maintain them at speed. The architecture has to be holistic from the first design decision.

Pillar 1 - Data Strategy & Architecture

Most organisations already have a data strategy document. The problem is that it was written once, presented to a steering committee, and has not been meaningfully revisited since. A document is not a strategy. A strategy is a living investment roadmap - one that connects today's architectural decisions to tomorrow's AI ambitions, and that evolves as the business, the technology, and the competitive landscape change around it.

Getting the architecture right at the outset is the highest-leverage decision an organisation can make in its AI programme. Architectural debt compounds faster than almost any other form of technical debt - because every AI use case, every model, and every downstream system built on a poorly designed foundation inherits its constraints. Rebuilding architecture mid-programme is not just expensive. It is the single most common reason AI transformation timelines slip by years rather than months.

A well-designed data strategy and architecture investment roadmap addresses four things that a one-off document typically does not: how the architecture will evolve as AI use cases mature; where investment needs to be sequenced to unlock the next layer of capability; what the organisation's data estate will need to look like in three years, not just today; and how architectural decisions made now will affect the cost, speed, and risk profile of every AI initiative that follows.

For C-suite leaders, the right question is not "do we have a data strategy?"
It is "does our data strategy tell us what to invest in next, and why?"

Pillar 2 - Cloud & Data Platforms

The cloud conversation in 2026 is no longer about whether to move to the cloud. It is about where data lives, who controls it, and what it costs to run AI at scale on top of it. For most enterprises, the answer is not a single cloud - it is a deliberate hybrid architecture that balances performance, cost, and control.

Hybrid cloud has matured from a transitional compromise into a strategic choice. Organisations operating across multiple geographies, regulated industries, or complex data sharing arrangements are increasingly finding that keeping certain data on-premise - or within specific cloud regions - is not a limitation. It is a requirement. GDPR, the EU AI Act, and a growing body of national data sovereignty legislation mean that where data resides is now a legal and regulatory question, not just an infrastructure one. Organisations that did not design for this are now retrofitting compliance into architectures that were never built to support it.

Hybrid cloud design is increasingly a requirement rather than an option. Organisations that put everything on public cloud are running into cost challenges as usage scales. Leading organisations are building across cloud for elasticity, on-premises for consistency and data sovereignty, and edge for latency-critical applications. The data engineering platform needs to operate coherently across all three - with unified governance, consistent data contracts, and seamless orchestration regardless of where compute actually runs.

The cost dimension is equally significant. Unconstrained cloud adoption - moving everything to a single hyperscaler and scaling compute on demand - made sense when AI workloads were experimental. At production scale, the economics change sharply. Egress fees, storage costs, and the compute demands of running large models continuously can turn a promising AI business case into a budget problem within a single quarter. Hybrid architectures give organisations the ability to run the right workload in the right environment - keeping sensitive or high-volume data where it is cost-effective and compliant to do so, while leveraging cloud elasticity where it genuinely adds value.

For C-suite leaders, the platform question is not just "are we in the cloud?" It is "do we have the architectural control to manage what our AI programme will cost – and what our regulators will require – two years from now?"

Pillar 3 - Governance & Compliance by Design

The EU AI Act is now in force. For any enterprise operating in or selling into European markets, this is not a future consideration – it is a current compliance obligation with real penalties attached. Combined with GDPR, emerging national AI legislation, and the growing body of case law around automated decision-making, the regulatory environment for enterprise AI has shifted from ambiguous to enforceable. Organisations that have not designed governance into their data engineering infrastructure are already carrying exposure they may not yet have quantified.

The instinct in most organisations is to treat governance as a layer added on top of existing infrastructure – a set of policies, processes, and audit trails bolted onto pipelines that were never designed to support them. This approach has a fundamental weakness: it relies on human intervention at every point where a governance decision needs to be made. At the data volumes and decision speeds that production AI operates at, manual governance does not scale. It breaks – quietly, and usually at the worst possible moment.

Governance by design means something different. It means access controls, privacy rules, lineage tracking, and regulatory requirements are embedded into the data platform itself – enforced automatically, continuously, and without requiring an engineer or compliance officer to intervene at each data call. Every dataset knows where it came from, what rules apply to it, and who is permitted to use it for what purpose. Every AI output can be traced back to the data that produced it.

This matters for two reasons that resonate at board level. The first is regulatory defence – when a regulator or a court asks how a particular AI decision was made, the answer needs to be retrievable, accurate, and fast. The second is operational trust – AI systems that cannot explain their outputs, or whose data provenance cannot be verified, will not be adopted at scale by the business units that need to rely on them. Governance is not just about avoiding fines. It is about building AI that the organisation can actually stand behind.
For C-suite leaders, the question is not whether governance is important. It is whether your data engineering infrastructure enforces it automatically – or whether your compliance posture depends on people remembering to follow a process

Pillar 4 - AI & Data Validation

The quality of what an AI system produces is bounded by the quality of the data it receives – both during training and at inference time. That makes data validation a direct determinant of AI reliability, commercial value, and risk, not just a data engineering concern.

Rigorous validation in a platform-grade environment goes well beyond periodic quality reviews. It means automated schema validation at ingestion, statistical monitoring to detect drift between training and production data, anomaly detection to catch unexpected patterns before they reach AI decision systems, and throughput benchmarking to keep pipelines within the operational bounds each use case demands.

The 1-10-100 rule of data quality hits harder with AI systems. An error caught at ingestion costs a fraction of one caught during model evaluation, which costs a fraction of one that surfaces as a wrong output in production. For AI making real-time commercial decisions – dynamic pricing, credit assessment, clinical recommendations, supply chain calls – a quality failure reaching the output layer can be catastrophic. Validation built into the pipeline isn't a quality overhead. It's a core risk management function.

"AI-first organisations that report mature, well-established data and governance frameworks are more than twice as likely to achieve significant AI returns as those without structured data foundations."

– IBM Institute for Business Value, 2025

Pillar 5 - **DevOps for Data & AI**

Every COO understands the cost of a production line that stops. In data and AI, that production line is the pipeline - and most organisations have no early warning system when it starts to fail. By the time a broken pipeline surfaces - through a dashboard that stops updating, a model that starts producing unreliable outputs, or an AI-powered process that quietly degrades - the downstream damage is already done. DataOps is the discipline that prevents this. It brings the same logic that transformed manufacturing efficiency to data and AI engineering: quality built into every stage of the process, not inspected at the end of it.

Without it, the alternative is what most organisations are currently living with - data pipelines that are built once, tested informally, and deployed with fingers crossed. Changes are made manually, often by the person who originally built the pipeline and whose knowledge lives nowhere else. When something breaks, the diagnosis takes hours. When a new AI model needs different data, the engineering queue backs up. The result is a data function that is permanently reactive, permanently behind, and permanently expensive to run.

DataOps - and its AI-specific counterpart, MLOps - introduces the discipline of continuous integration and continuous delivery to data and model pipelines. In plain terms, this means every change to a data pipeline or AI model goes through an automated process of testing, validation, and controlled deployment before it reaches production. Quality gates replace manual checks. Rollbacks are automated rather than surgical. Monitoring is continuous rather than reactive.

The business outcome is not abstract. Organisations with mature DataOps practices deploy data changes significantly faster, with fewer production incidents and lower remediation costs. DataOps-guided engineering teams are projected to be ten times more productive than those operating without platform-level practices. For a COO focused on operational efficiency and cost control, that productivity differential is the equivalent of getting ten engineers worth of output from one - without the hiring cost, the onboarding time, or the retention risk.

For C-suite leaders, the right question is not "do our engineers follow good practices?" It is "does our data engineering infrastructure enforce quality automatically - so that the organisation is not one bad deployment away from an AI outage?"

Section 5

What Platform-Grade Engineering Delivers

In 2026, the AI arms race has produced an unexpected levelling effect. Every enterprise now has access to the same frontier models, the same hyperscaler infrastructure, and broadly the same tooling ecosystem. The technology is no longer the differentiator. The differentiator is the data loop - how fast, how reliably, and how trustworthily an organisation can move from raw data to a decision that creates value. Platform-grade data engineering is what determines that speed. Everything else is table stakes.

AI Initiatives That Reach Production

The most direct return on data engineering platform investment is the ability to actually move AI from pilot to production. Deloitte's 2026 State of AI in the Enterprise report found that only 25% of organisations have moved 40% or more of their AI pilots into production - but that 54% expect to reach that level within three to six months. The organisations that will get there are not those with access to better models - every competitor has the same models. They are the ones whose data engineering infrastructure can handle the demands of production AI: consistent pipelines, governed data, real-time access, and quality enforcement that holds up at scale.

The organisations that will not get there are the ones still working pipeline by pipeline, without the platform architecture that makes AI deployment safe and governable. As AI-native competitors build on their platform advantage quarter by quarter, the gap will widen - and the cost of catching up will grow with it.

Decision Velocity That Compounds

When every enterprise is running the same LLMs, the competitive question is no longer which model you are using. It is how quickly and reliably your data reaches it. A pricing team that can respond to market movements in minutes rather than hours. A risk team that sees exposure in real time rather than the next morning. A customer experience team that personalises at the point of interaction rather than in the next campaign cycle. None of this is about model sophistication. All of it is about data infrastructure.

This compounds strategically. Every cycle where a platform-grade organisation makes a faster, better-informed decision than a pipeline-grade competitor is a cycle where the gap widens. In markets where pricing accuracy, supply chain responsiveness, or credit decisioning speed is a genuine competitive differentiator, data engineering quality directly determines market position. The model is the commodity. The data loop is the moat.

Cost Efficiency That Scales as Workloads Grow

The platform approach requires more upfront architectural investment - and it consistently delivers lower total cost of ownership over time. Modern cloud-native data stack adopters achieve 50% lower total cost of ownership compared to fragmented pipeline environments. Platform components get reused across use cases, governance frameworks are inherited rather than rebuilt, and operational costs are kept in check through automated monitoring rather than manual firefighting.

The alternative - accumulating technical debt through pipeline proliferation - gets exponentially more expensive. Each new pipeline brings new integration points, new failure modes, and new maintenance overhead. Each AI initiative built on fragmented infrastructure inherits the cumulative quality risk of every upstream pipeline it depends on. And as AI workloads grow, those costs scale with them. The platform approach reverses that dynamic - costs stabilise as capability expands, rather than growing in lockstep with it.

This cost stability also has a talent dimension that C-suite leaders increasingly recognise. The best data engineers do not want to spend their careers firefighting brittle pipelines and rebuilding integrations that should have been standardised years ago. They leave for organisations that offer platform-grade engineering - modern tooling, architectural integrity, and problems worth solving. In a market where a senior data engineer is scarce, expensive, and takes months to replace, attrition in this function is not an HR problem. It is a delivery risk. Organisations that invest in platform-grade infrastructure are, among other things, investing in an environment that retains the people they cannot afford to lose

KEY STATISTICS

10x
productivity improvement for DataOps-guided engineering teams vs. non-platform teams
(Industry analysis, 2026)

50%
lower total cost of ownership for modern cloud-native platform adopters vs. fragmented pipeline environments
(Industry benchmarks, 2025)

Regulatory Confidence as AI Scrutiny Increases

For organisations in regulated sectors - financial services, healthcare, energy, insurance - data governance is not a strategic choice. It is a legal requirement, and it is tightening as AI deployment expands. Here again, the LLM playing field is level - every competitor can access the same models. What is not level is the ability to deploy those models in a way that is demonstrably compliant. The EU AI Act's requirements for data provenance and quality, combined with existing GDPR obligations and sector-specific regulations, mean that organisations without embedded governance face growing exposure as they scale AI.

A data engineering platform with governance built into the architecture is not just a technical advantage in this environment. It is a prerequisite for deploying AI responsibly - and a structural advantage over competitors who are still trying to bolt compliance onto infrastructure that was never designed to support it.

Section 6

Where to Start: A Practical Roadmap

The most common and most costly mistake organisations make at the start of a data engineering transformation is reaching for technology before they have defined their architecture. A new cloud platform, a modern data lakehouse, a best-in-class orchestration tool - none of these investments will deliver their promised return if the architectural foundation underneath them has not been properly designed. In practice, this is how organisations end up with sophisticated tooling that does not fit their data landscape, expensive licences for platforms that cannot be fully adopted, and integration complexity that creates more problems than the tools were brought in to solve.

Architecture before technology is not a philosophical preference. It is a commercial discipline. Getting the architecture right first means understanding what the organisation's data estate actually looks like today, where the gaps are, what production AI will demand of the infrastructure in two to three years, and what sequencing of investment will unlock the most value at each stage. Only once that picture is clear does technology selection become a decision with a defensible answer - because the architecture defines what the technology needs to do, rather than the other way around.

The practical implication for C-suite leaders is straightforward. Vendors will always have a solution to sell. The question is not whether their solution works in isolation - most do. The question is whether it fits the specific architectural context of your organisation, integrates cleanly with what you already have, and is the right investment at this stage of your data engineering maturity. Without an architectural blueprint, there is no basis on which to answer that question confidently. With one, technology decisions become faster, cheaper, and significantly less likely to require expensive reversal twelve months later.

Phase 1 - Assess and Architect

Before any technology is selected or any pipeline is built, the organisation needs a clear-eyed assessment of its current data estate - what exists, what is working, what is failing silently, and what the gap is between current infrastructure and what production AI will require. This assessment is the foundation on which every subsequent investment decision is made. It is also the stage that most organisations either skip entirely or compress into a few weeks of workshops that produce a slide deck rather than an actionable architectural blueprint.

A genuine architectural assessment covers data sources and their quality characteristics, pipeline dependencies and fragility points, governance gaps and compliance exposure, and the latency and freshness requirements of the AI use cases the organisation is prioritising. The output is not a technology recommendation. It is an architectural design and an investment roadmap - a sequenced plan that shows what needs to be built in what order, and why.

Phase 2 - Stabilise and Govern

With the architecture defined, the next priority is stabilising the data foundation that AI will run on. This means addressing the most critical quality and governance gaps identified in the assessment - not comprehensively, but strategically, focusing on the data domains that matter most to the AI use cases in the pipeline. Governance frameworks get embedded at this stage, not added later. Access controls, lineage tracking, and data contracts are built into the infrastructure from the outset, so that every workload built on top of them inherits compliance by default rather than requiring it to be retrofitted.

Phase 3 - Modernise and Scale

With a stable, governed foundation in place, the organisation can begin modernising its broader data infrastructure with confidence - migrating legacy pipelines to platform-grade architecture, introducing real-time data capabilities where AI use cases require them, and expanding self-service access so that data science and AI teams can move faster without creating engineering bottlenecks. Technology selection happens here, informed by the architectural blueprint defined in Phase 1 rather than by vendor relationships or market trends.

Phase 4 - Optimise and Compound

The final phase is where platform thinking delivers its compounding return. As new AI use cases are built on the platform, they inherit the governance, quality, and reliability already embedded in the infrastructure. Engineering effort shifts from building and maintaining bespoke pipelines to extending a platform that gets more valuable with every workload added to it. Decision velocity increases. Costs stabilise. And the organisation builds a data capability that competitors working pipeline by pipeline will find increasingly difficult to close the gap on.

Merit's Approach to Data Engineering Platform Design

Merit's data engineering practice is built on a simple conviction: data infrastructure is a strategic asset, not a technical utility. We design AI-ready data platforms from the architecture up - defining governance models before selecting technologies, embedding quality enforcement into pipelines rather than appending it, and applying DataOps discipline from day one. Our clients don't inherit technical debt from their data platform. They inherit competitive advantage that compounds with every use case built on it.

Conclusion

The question facing every enterprise leader with an AI agenda in 2026 is not whether to invest in AI. It is whether the data engineering infrastructure beneath that investment can support production at the scale the ambition requires.

The pilot-to-production gap frustrating AI investment across every sector is a data engineering problem - not a model problem, not a use-case problem, not a skills problem. It is an infrastructure architecture problem. One that pipeline thinking cannot solve. One that platform thinking can.

The organisations that make the right architectural decisions in 2026 will not simply be ahead in 2028. They will have spent two years compounding an advantage that rivals cannot close quickly - because platform-grade data infrastructure is not something that can be replicated in a quarter. It is built through sustained, sequenced investment, and the organisations that started early will have the production AI, the decision velocity, and the regulatory confidence to show for it.

The ones that did not will still be in the gap - managing more pipelines, carrying more technical debt, and watching the distance grow. 2026 is not the moment to plan the transition from pipelines to platforms. It is the moment to make it.

Ready to move from pipelines to platforms?

Merit's data engineering teams design and build AI-ready data platforms for organisations in energy, financial services, maritime, healthcare, and beyond. We work from strategy and architecture through to production - ensuring the infrastructure beneath your AI investment is built to scale, govern, and deliver.

meritdata-tech.com

About Merit Data & Technology

Merit Data and Technology is part of Merit Group PLC. For over 20 years, we have been helping enterprises design and build data infrastructure that performs under real-world conditions - not just in the proof of concept, but in production, at scale, and under regulatory scrutiny.

Two decades in this field means we have seen technology cycles come and go. We have worked with organisations that got the architectural decisions right early and compounded the advantage - and with organisations that deferred those decisions and paid a steep price to recover. That experience shapes how we approach every engagement - with the rigour, the realism, and the long-term perspective that high-stakes data transformation decisions deserve.

For C-suite leaders who need more than a vendor - who need a partner with the track record to be trusted with decisions that will define their organisation's AI trajectory for years to come - Merit brings 20 years of delivery credibility to the table.

Sources

1. Deloitte AI Institute (2026). State of AI in the Enterprise 2026. Survey of 3,235 senior leaders across 24 countries, August-September 2025. deloitte.com/global/en/issues/generative-ai/state-of-ai-in-enterprise.html

2. Deloitte AI Institute (2026). State of AI in the Enterprise 2026. Finding: only 25% of respondents have moved 40%+ of AI pilots into production. Ibid.

3. IBM Institute for Business Value (2025). 2025 CEO Study. Finding: only 16% of AI initiatives have successfully scaled across the enterprise. ibm.com/thought-leadership/institute-business-value

4. BCG (2025). Enterprise AI Adoption Research. Finding: 74% of companies struggle to scale AI value due to data governance and accessibility issues. bcg.com

5. IBM Institute for Business Value (2026). Biggest Data Trends 2026. Finding: enterprises typically have up to 90% of their data locked in unstructured silos. ibm.com/think/news/biggest-data-trends-2026

6. IDC (2025). AI Readiness and Data Management Survey. Finding: 83% of enterprises have shifted data management priorities in response to AI demands. Cited in: IBM Think Insights, November 2025.

7. Integrate.io / Industry Analysis (2025). Data Pipeline Efficiency Statistics. Finding: 31% of organisations report revenue loss attributable to data lag or pipeline downtime. integrate.io/blog/data-pipeline-efficiency-statistics

8. IBM Institute for Business Value (2024). AI-Ready Enterprise Data Survey. Finding: only 29% of technology leaders strongly agree their enterprise data meets the standards required to scale generative AI. ibm.com/thought-leadership/institute-business-value

9. DataToBiz / Industry Analysis (2025). Top Data Engineering Trends. Finding: organisations manage 5-7+ specialised data tools on average; 70% of data leaders cite stack complexity as a significant challenge. datatobiz.com/blog/top-data-engineering-trends

10. Grand View Research (2024). Cloud Data Pipeline Market Analysis. Finding: 71.18% of data pipeline tools are now cloud-based deployments. Cited in: Integrate.io Data Pipeline Efficiency Statistics, 2025.

11. European Union (2024). Regulation (EU) 2024/1689 - Artificial Intelligence Act. Entered into force August 2024; two-year transition period for providers of high-risk AI systems. eur-lex.europa.eu

12. Industry Analysis (2026). DataOps Productivity Benchmarks. Finding: data engineering teams guided by DataOps practices projected to be 10x more productive than teams without platform-level practices. datatobiz.com/blog/top-data-engineering-trends

13. Deloitte AI Institute (2026). State of AI in the Enterprise 2026. Finding: 54% of organisations expect to move 40%+ of AI pilots to production within three to six months. Ibid.

14. Industry Benchmarks (2025). Modern data stack adopters achieve 50% lower total cost of ownership vs. fragmented pipeline environments. Cited in: Integrate.io Data Transformation Challenge Statistics, January 2026.