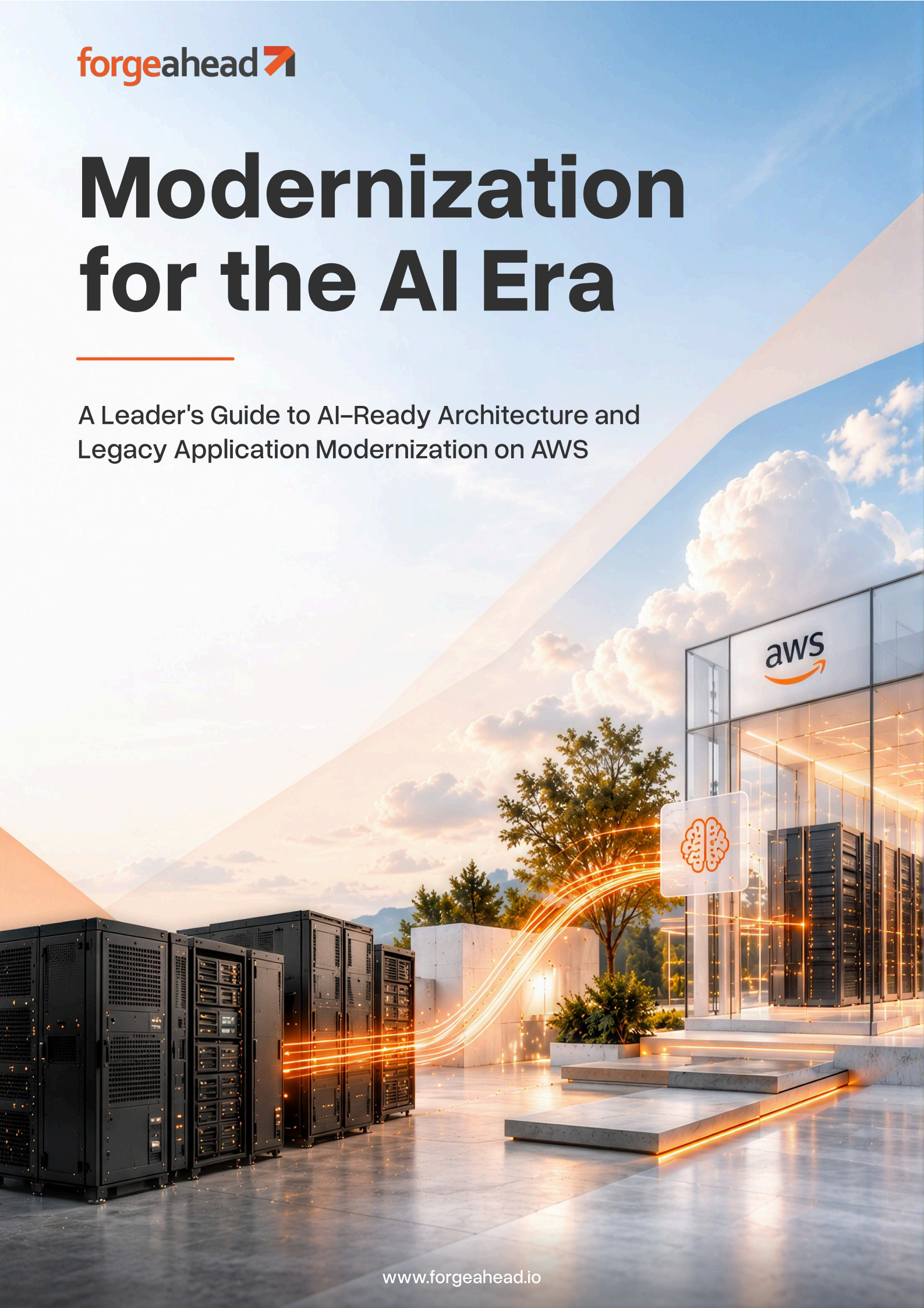
forgeahead

Modernization for the AI Era

A Leader's Guide to AI-Ready Architecture and Legacy Application Modernization on AWS



www.forgeahead.io

Table of Contents

Executive Summary 03

Why Legacy Applications Stall Enterprise AI 04

What AI-Ready Architecture Looks Like on AWS 06

How to Modernize Legacy Applications for Enterprise AI 09

Building the Data Foundation AI Depends On 11

Architecture Patterns That Work 13

Measuring What Matters: Readiness, ROI, and KPIs 15

Choosing the Right Modernization Partner 17

Get Started 17

References 18

Executive Summary

The AI pilot worked. The demo impressed the board. Then someone asked how it would connect to the system of record, and the project went quiet.

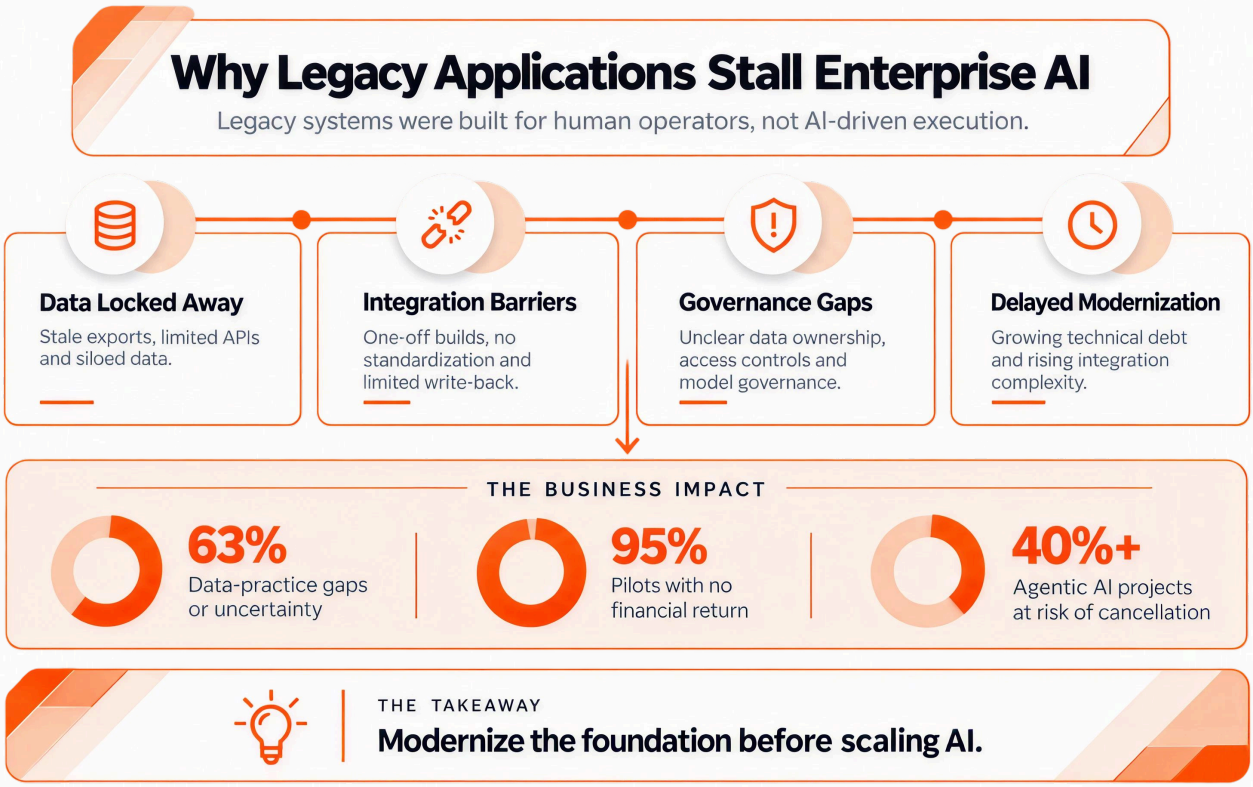
That pattern repeats across enterprises, and the research backs it up. Deloitte's 2026 State of AI in the Enterprise report found that only 25% of organizations have moved at least 40% of their AI pilots into production, with technical infrastructure readiness at 43% and talent readiness at 20%. Gartner expects organizations to abandon 60% of AI projects that lack AI-ready data. The constraint usually lies in the applications and data underneath the models.

CTOs and heads of engineering and data who have already approved an AI budget now face a harder question, which is what has to change in the existing application estate for that budget to produce results. This guide covers why legacy applications stall enterprise AI, what AI-ready architecture involves on AWS, a phased approach to legacy application modernization for AI that avoids a big-bang rewrite, how to build the data foundation, and how to measure progress.



Why Legacy Applications Stall Enterprise AI

Most legacy applications were designed around a human operator. Data is stored in tables shaped for transactions, business rules are embedded in application code, and the only ways in are a user interface and a nightly batch export. An AI system needs a different kind of access. It needs governed, current data through an API, enough context to interpret what a record means, and a path to act on its output without anyone re-keying it.



Where the cost actually shows up

The cost rarely appears as a single line item. It appears as pilots built on exported spreadsheets because live data was off limits, and as one-off integrations rebuilt for every new use case. Add-ons that can read but never write back are common too. Security reviews stretch because nobody can show exactly which data a model touched. Each workaround is reasonable alone, and together they explain why so few pilots reach production.

The research points the same direction. Gartner's survey found that 63% of organizations either lack the right data management practices for AI or are unsure whether they have them, and Gartner predicts that through 2026 organizations will abandon 60% of AI projects unsupported by AI-ready data. MIT's NANDA study reviewed over 300 enterprise generative AI initiatives and found that 95% of the pilots showed no financial return on the P&L. A pilot with no path into production systems has nothing to show on a P&L.

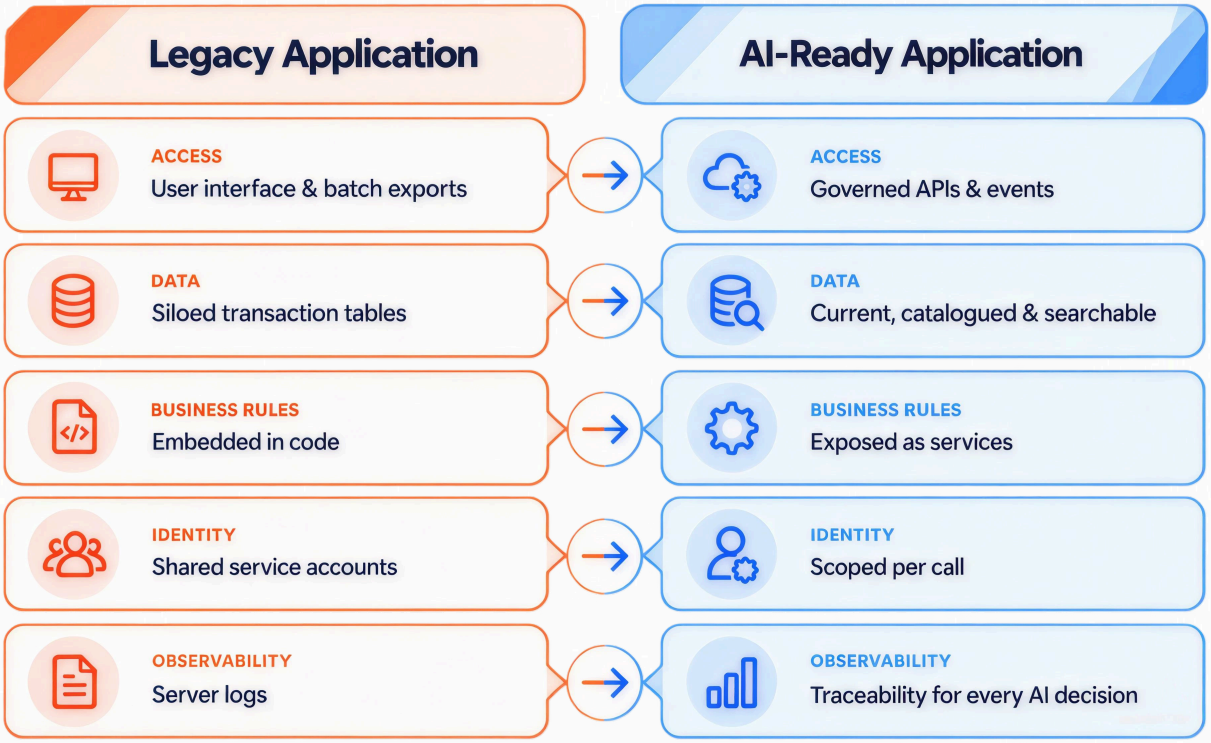
Why "eventually" is expensive

Spending is moving toward custom AI applications. Gartner's September 2026 forecast puts worldwide AI spending at $2.7 trillion for the year, up 49.5%, and raised its 2026 growth outlook for AI application development platforms to 39% as enterprises build applications tailored to their own needs. Applications like that need integration points the existing estate may not have. Every year spent postponing modernization adds integrations built around the old limits, and each one becomes another thing to unwind later.

Gartner also expects over 40% of agentic AI projects to be canceled by the end of 2027. Escalating costs and inadequate risk controls are among the causes Gartner names. Foundations decide which side of that figure a program lands on.

What AI-Ready Architecture Looks Like on AWS

AI-ready applications let other systems ask them to do things. A claims platform exposes "check coverage" and "update status" as secured API calls. A contract archive exposes search and retrieval. Business rules that used to be embedded in code become services with defined inputs and outputs, so a model or an agent can call them and a person can audit what happened.



The layers that matter

AI-ready architecture on AWS tends to follow the same layers regardless of industry. The table below maps each layer to its purpose and the AWS services that usually deliver it.

Layer	Purpose	Typical AWS services
Integration	Expose legacy capabilities as secured APIs and events	Amazon API Gateway, Amazon EventBridge, AWS Step Functions
Data	Store governed, current, searchable data for models to use	Amazon S3, AWS Lake Formation, AWS Glue, Amazon Aurora, Amazon OpenSearch Service
Model and agent	Run foundation models and coordinate multi-step agent work	Amazon Bedrock, Amazon Bedrock AgentCore
Compute	Run application logic matched to workload shape	AWS Lambda, Amazon ECS, AWS Fargate
Governance and observability	Control access, protect data, and trace every AI decision	AWS IAM, AWS KMS, Amazon Bedrock Guardrails, Amazon CloudWatch, AWS CloudTrail

Cloud modernization for AI is not a standard migration

A rehosted application on EC2 has the same constraints it had in the data center, so it is no more ready for AI than before. Cloud modernization for AI asks a different question of each application, namely which capabilities a model or agent will need to call and what data those capabilities depend on. The answers decide how far the modernization goes. Some applications need only an API layer. Others need a database replatformed or a capability extracted into its own service.

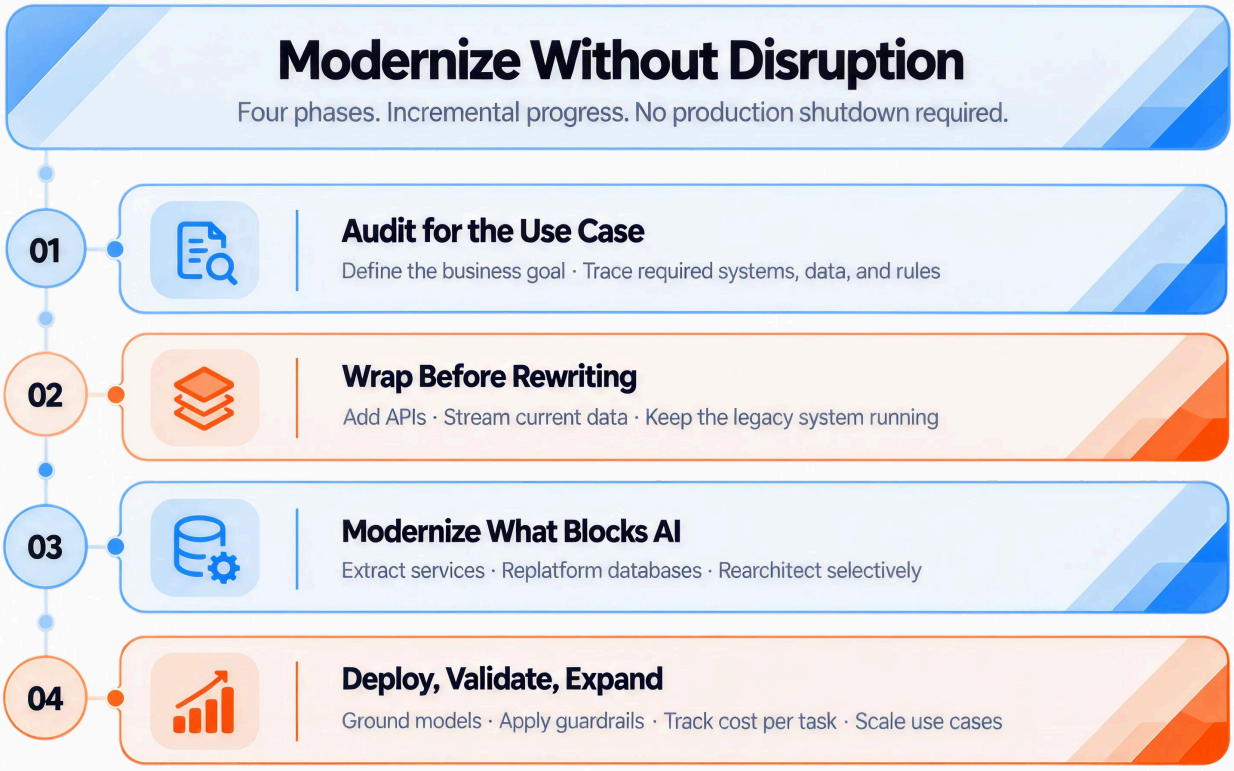
AWS has been lowering the effort of the code-level work. AWS reported in May 2026 that, in its first year, AWS Transform helped thousands of customers migrate hundreds of thousands of servers and process over 4.5 billion lines of code. In August 2026 it made continuous modernization generally available for analyzing and remediating technical debt across code repositories. Vendor-reported figures are best treated as indicative, so test the tooling on one codebase before building a program around it.

Enterprise AI infrastructure is mostly about integration

Gartner's May 2026 forecast notes that AI infrastructure is the largest segment of AI spending, at over 45% of the total, and that vendors and hyperscalers drive it. Most enterprises consume that capacity through managed services such as Amazon Bedrock without buying servers, so their enterprise AI infrastructure decisions look different. They involve identity and data access around models that someone else operates, along with the cost controls and audit trails that make those models safe to run.

How to Modernize Legacy Applications for Enterprise AI

How to modernize legacy applications for enterprise AI comes down to sequencing. Four phases work for most enterprises, and none of them require taking a production system offline.



Phase 1: Start with the use case, then audit the application

Pick one AI use case tied to a business number, such as turnaround time or cost per transaction. Trace which systems, data, and rules it needs. That trace becomes the audit scope, which keeps discovery limited to what the use case depends on. Check data readiness against the use case at the same time, since Gartner's guidance starts with aligning data sources to specific AI use cases and testing data quality continuously.

Phase 2: Wrap before you rewrite

Put an API layer in front of the legacy system and expose read paths first. Stream changes out of the database through change data capture so AI components work from current data without querying production tables directly. AWS Database Migration Service supports ongoing replication for this purpose. The legacy system keeps running as the system of record, and the AI layer builds on top of it.

Phase 3: Modernize the parts that block AI

Some capabilities cannot be wrapped because the code behind them is too tangled or too slow. Extract those into services, replatform a self-managed database onto Aurora where the data layer limits retrieval, and rearchitect only the workflows an AI use case needs. Match compute to workload shape.

Phase 4: Add AI in stages and expand

Introduce models with grounding and guardrails. Retrieval over approved data keeps answers tied to verifiable sources, human review handles low-confidence cases, and Amazon Bedrock Guardrails limits what a model can say or do. Track cost per completed task from the first week. Once the first use case holds in production, onboard the second onto the same integration layer, which is where cost per use case starts to fall.

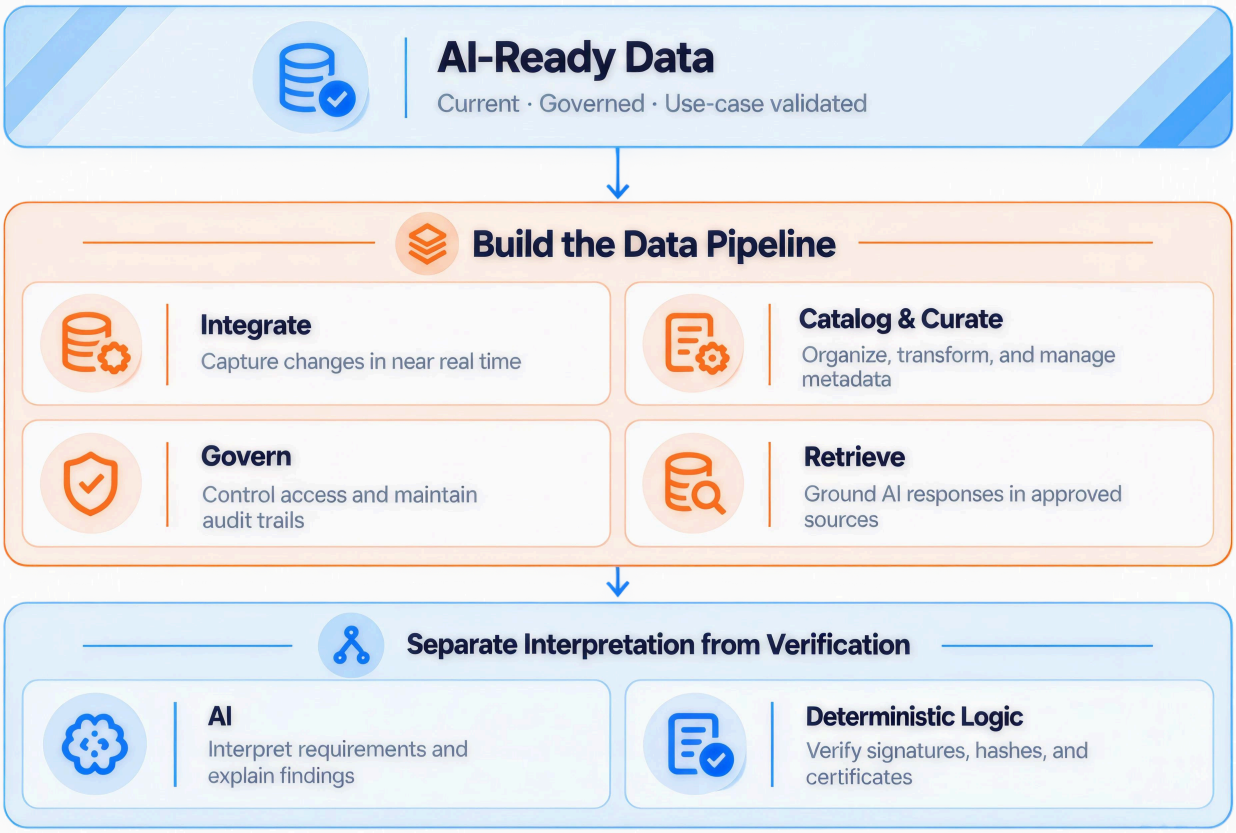
What makes this approach work

Nothing here requires a freeze on feature work or a single high-stakes cutover. Each phase ends with a working result, so leadership sees progress while the estate underneath improves.

Building the Data Foundation AI Depends On

Data that serves reporting well can still fail an AI use case. A day of lag and inconsistent field names are tolerable in a report, where a person knows which column to trust. A model has no such context. AI-ready data is current and governed, with quality tested against the use case it serves.

Gartner notes that organizations with basic or manual metadata management will find it hard to make data AI-ready. Maturing metadata practices is the first practical step.



The components on AWS

Amazon S3 holds raw and curated data, AWS Lake Formation applies fine-grained access control, and AWS Glue catalogs and transforms it. Change data capture feeds events into Amazon EventBridge or Amazon Kinesis so downstream systems see updates quickly. Vector search through Amazon OpenSearch Service or Aurora PostgreSQL gives retrieval something to query. Governance stays attached to the data. Lake Formation permissions and CloudTrail records follow every access, which matters when an auditor asks what a model saw.

Retrieval grounded in the system of record

Retrieval-augmented generation (RAG) grounds AI responses in approved enterprise sources, such as historical proposals, contracts and compliance documents. This helps reduce unsupported outputs and ensures responses draw on relevant, verifiable information, particularly in high-stakes workflows where missing a critical requirement can have significant consequences.

Deterministic logic remains equally important. In compliance workflows, rule-based checks can verify digital signatures, cryptographic hashes, certificate chains and revocation status, while AI interprets complex requirements and explains findings. Code handles verifiable checks; AI supports interpretation. This separation improves accuracy, traceability and trust.

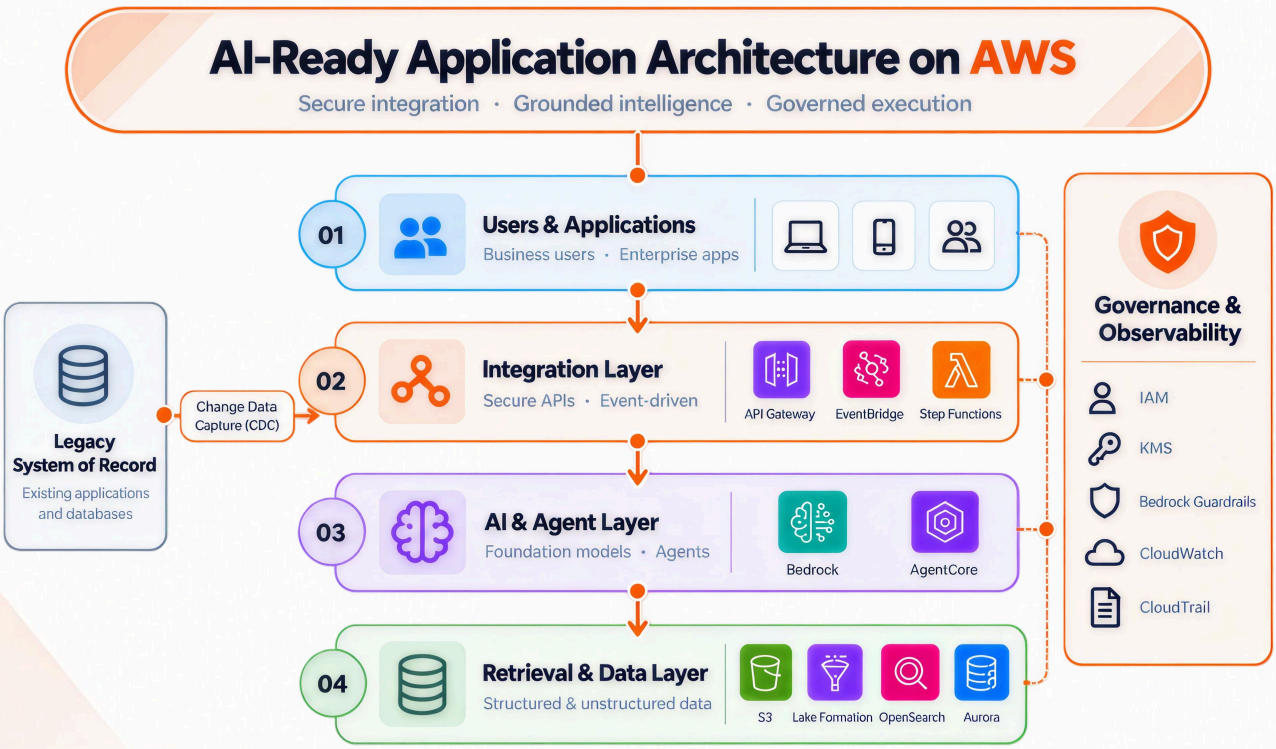
Enterprise AI is a data problem before it is a model problem

Worth saying plainly. Most disappointing AI results trace back to data that was stale or hard to reach. Organizations that fix the pipeline first get more from every model they adopt afterward, and model choice becomes easier to change.

Architecture Patterns That Work

A modernized platform for enterprise AI on AWS follows a similar pattern regardless of company size. The legacy system remains the system of record. Change data capture streams updates out of it, and Amazon API Gateway exposes its capabilities as secured APIs. Amazon S3 and Lake Formation hold governed data, and OpenSearch provides retrieval.

Amazon Bedrock supplies the models, with an abstraction layer so workloads can route to different models as needs change. Agents coordinate through the Strands Agents SDK or Amazon Bedrock AgentCore, and AWS Step Functions tracks state for fixed-order workflows. AWS recently added runtime instances to AgentCore for agents that need persistent compute and sessions lasting up to 14 days. Compute runs on Lambda for event-driven work and on ECS with Fargate for heavier processing. CloudWatch and CloudTrail record what happened.



Security, governance, and cost belong in the architecture

Enterprise AI infrastructure carries sensitive data and autonomous behavior, so access controls and audit trails cannot wait for a later release. Each agent and function should run under its own narrowly scoped IAM role, with KMS encryption and private networking through VPC endpoints. Customer data should stay out of model training. In the compliance platform, submitted documents were processed in memory and never persisted.

Agent governance deserves its own attention. Deloitte found that close to three-quarters of companies plan to deploy agentic AI within two years, while only 21% report a mature governance model for it. AWS has begun addressing this directly. In August 2026 it open-sourced Dogwood, a policy language for AI agents, and Bedrock AgentCore added temporal policies that evaluate an agent's history of actions within a session.

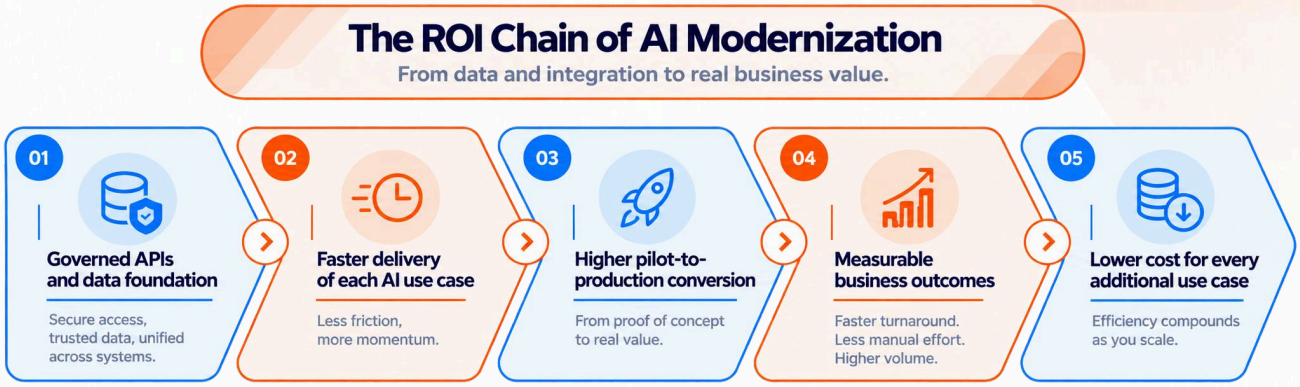
Built to extend

The payoff of this architecture is the second and third use case. Once the integration layer, data foundation, and governance exist, a new AI use case builds against APIs and policies that are already in place. The RFQ platform's LLM abstraction layer is a concrete example. It routes tasks to different Bedrock models, so the platform can adopt a new model without changes to the systems upstream. Models will keep changing, and an architecture that treats them as replaceable absorbs that change at low cost.

Measuring What Matters: Readiness, ROI, and KPIs

A modernization program for AI needs measures at four levels, each with a baseline set before work begins.

Level	Metric	What it tells you
Readiness	Share of priority capabilities exposed through governed APIs	Whether AI use cases have something to call
Readiness	Share of AI-relevant data with documented lineage and quality checks	Whether data meets the AI-ready standard
Delivery	Deployment frequency, lead time for changes, change failure rate, time to restore service	Whether change has become cheaper
AI delivery	Pilot-to-production conversion rate and time from approved use case to production	Whether the platform shortens the path to value
AI delivery	Cost per completed task and human override rate	Whether agents are efficient and trusted
Business	Turnaround time, manual effort, and volume handled per person	Whether the architecture moves a business number



Pilot-to-production conversion deserves leadership attention. Deloitte found that only 25% of organizations have converted at least 40% of their AI pilots into production, so a rising conversion rate is direct evidence that the foundation is working. Cost per completed task is a better signal than cost per token, since it counts retries and failed attempts that a per-token figure hides.

Choosing the Right Modernization Partner

AI modernization programs usually fail on integration with existing systems, on timelines built around a single big cutover, or when a partner delivers a demo and leaves before anyone has to operate it. MIT's study found that organizations working with external partners reached deployment about 67% of the time, against 33% for internal builds. Look for AWS-specific engineering depth, experience connecting AI to existing systems of record, incremental delivery, and accountability for running the platform after launch.

Get Started

Preparing legacy applications for enterprise AI starts with understanding which systems your AI initiatives will depend on, where their data lives, what integration gaps exist, and which modernization steps can be taken without disrupting daily operations.

Schedule an AWS application modernization assessment with Forgeahead.



forgeahead

References

- https://www.deloitte.com/cz-sk/en/services/consulting/research/the-state-of-ai-in-the-enterprise.html
- https://www.gartner.com/en/newsroom/press-releases/2025-02-26-lack-of-ai-ready-data-puts-ai-projects-at-risk
- https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/
- https://www.gartner.com/en/newsroom/press-releases/2026-09-16-gartner-forecasts-worldwide-ai-spending-to-grow-49-point-5-percent-in-2026
- 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
- https://aws.amazon.com/blogs/migration-and-modernization/aws-transform-one-year-milestone/
- https://www.gartner.com/en/newsroom/press-releases/2026-05-19-gartner-forecasts-worldwide-ai-spending-to-grow-47-percent-in-2026
- https://aws.amazon.com/about-aws/whats-new/2026/08/aws-bedrock-agentcore-runtime-instances-generally-available/
- https://www.deloitte.com/us/en/insights/topics/emerging-technologies/ai-agents-scaling-faster.html
- https://www.deloitte.com/nl/en/about/press-room/state-of-ai-in-the-enterprise-2026.html
- https://aws.amazon.com/blogs/opensource/introducing-the-dogwood-local-engine-temporal-governance-for-agent-actions/