softserve | aws

FUTURE OF AI IS CLOUD

If you want your AI to take off, you've got to modernize your infrastructure, cloud, and on-prem alike. Cloud-native AI and platform engineering are the keys to speed, scalability, and innovation.

Ruslan Kusov, Cloud CoE Director, SoftServe
Ryan Peterson, WW Tech Leader, Modernization, AWS
Ayan Jain, WW Product Specialist, Modernization, AWS

Executive Summary

Cloud infrastructure is essential to enterprise AI success. As organizations accelerate AI adoption, they must modernize and upgrade legacy systems, shift workloads to cloud-native platforms, and embed automation and security. The goal is to continuously enhance applications, improve operational reliability, and integrate emerging technologies like AI. Without this foundation, most AI initiatives stall. In fact, nearly **90% of pilots never** reach production.

By 2030, companies that adopt an AI-first strategy, built on modern cloud platforms, will lead their industries. They will see lower costs, higher revenue, better security, and a competitive edge.

This report, a collaboration between AWS and SoftServe, outlines how leading companies close the gap between experimentation and business outcomes. It presents a practical roadmap for business leaders to align AI with goals, reduce technical debt, and scale innovation. Explore these approaches to find out which best fits your priorities.

Highlights

- **Why most AI projects fail**
- **Why cloud-native AI (CNAI) is the next step in enterprise architecture**
- **How platform engineering supports repeatable, secure AI deployment**
- **How clean, governed data unlocks AI value**
- **How to measure ROI, agility, and innovation velocity**
- **What to expect from agentic AI and how to prepare infrastructure now**
- **Strategies and achievements of other leaders**

Section 1: Why Most AI Projects Fail

AI is everywhere, except in production. A Forbes report said that nearly **9 out of 10 AI pilots** fail to reach full deployment. And it's not just because the technology is bad.

Misstep #1

Misaligned business value

Companies often begin with technological capabilities rather than business outcomes. Without clear ROI frameworks and measurable objectives, even technically successful pilots struggle to justify production scaling. AI that doesn't help a company earn more, save more, or compete better ends up in purgatory, stuck between experimentation and outcomes.

Misstep #2

Infrastructure gaps

Gartner predicts companies will cancel **over 40% of agentic AI** projects by 2027 due to higher costs, unclear business value, or inadequate risk controls. Why? Because getting agents to production requires fundamentally different infrastructure than traditional applications — one that's dynamic, secure, and built for autonomy. The deeper challenge is that AI agents are non-deterministic, meaning they can choose different paths to reach a goal. Meanwhile, your enterprise runs on predictable processes with compliance requirements, audit trails, and SLAs. Bridging these worlds is complex.

Most companies try to build from scratch. It's like building a modern web application, but having to first invent the web server, the database, and the programming language. The infrastructure simply doesn't exist today.

Companies must modernize and prepare for the demands of agentic systems. That starts in the cloud — but not just any cloud. The future of AI is built on modern, AI-ready platforms.

Booking.com moved from pilot to full-scale deployment because it built on a modern cloud foundation with AWS solutions (Amazon SageMaker and AWS Lambda) and aligned AI with business goals.

Learn How

Complexity

Scaling AI is a multidimensional problem. Complexity shows up across three areas.
Solve one, unlock the others.

Strategy:
Unclear ROI, fragmented stakeholders, misaligned priorities

Technology:
Infrastructure limitations, data silos, immature ML operations

Execution:
Cost management, scaling across hybrid/multi-cloud environments

Why is cloud essential for AI success?

Cloud provides the scalable, secure, and cost-efficient infrastructure AI needs to thrive. It supports dynamic workloads, enables faster experimentation, and integrates emerging technologies like agentic AI. By 2030, companies that adopt an AI-first approach will lead the way, backed by modern cloud foundations.

Meaning, every company should have a five-year plan to bring AI into its core operations, using cloud technologies to achieve this. This plan should focus on modernizing what's important and migrating what makes sense. Most companies are likely to face flat or even smaller budgets for 2026. A solid plan makes that smaller budget work for you, not against you.

We recommend four strategies to modernize:

- Adopt cloud-native AI (CNAI)
- Build strong data foundations
- Adopt platform engineering
- Use NVIDIA-accelerated solutions

Clients will see:

- Lower costs
- Higher revenue
- Stronger security
- Faster innovation
- Competitive edge

The data gap

AI needs company data to deliver value (and most of that data lives in the cloud). Companies that paused cloud migration to chase AI now realize that data modernization is a prerequisite for AI success.

SoftServe's study, **The Great Data Divide**, February 2025, found that only 22% of companies have achieved enterprise-wide Gen AI success. The rest are stuck because their data isn't ready.

The **AWS-commissioned Total Economic Impact study** by Forrester found that legacy data environments quietly drain resources, which causes delays, rework, and compliance risks. Moving data to the cloud isn't enough. You also need to modernize how it's integrated, scaled, and governed, especially the APIs and applications that feed your AI. This is crucial for AI workloads like retrieval-augmented generation (RAG), which rely on vectorized data to deliver context-aware results. Modern cloud-native tools (Amazon Bedrock Knowledge Bases and Amazon S3 Vector Databases) are already optimized for AI, which allows for faster, smarter data retrieval and inference at scale.

On the bright side, companies that invest in clean, structured, and well-governed data already see results: A 2024 Forrester global study commissioned by SoftServe found that **44% of companies report** new revenue streams from mature data management.

Tip

SoftServe outlines three pillars for a successful data and AI strategy:

1. Strong technology foundation
2. Embedded data governance
3. Clear business value alignment

Without all three, you'll build data ghost towns.

Data center investment and cloud spend soar

Investments in data centers are accelerating to support AI workloads.

Amazon has invested **$100 billion** in AI-focused data centers and infrastructure, marking its largest capital expenditure to date. Of that, **$30 billion** is earmarked for AI infrastructure across Pennsylvania and North Carolina, creating the technological backbone for the generative and agentic AI era.

This surge reflects a broader trend: Companies are making AI foundational. And that foundation is built on cloud infrastructure. Globally, the story is similar.

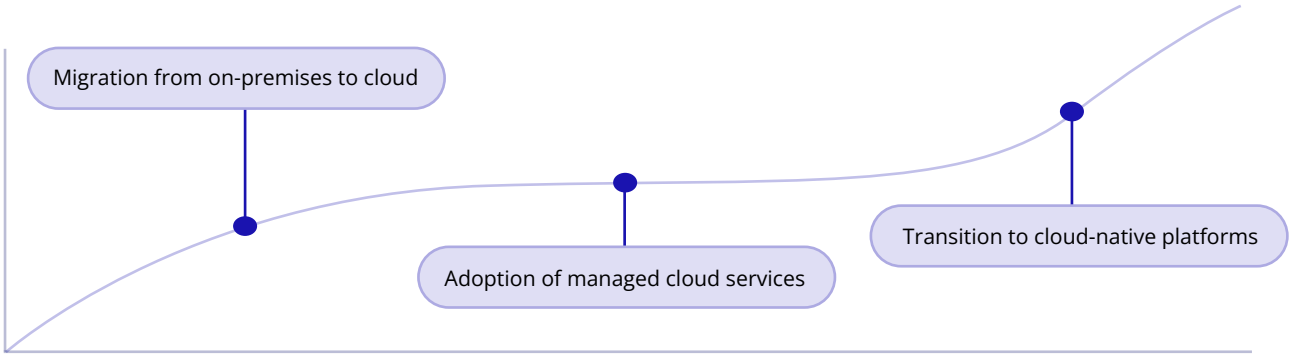
Datacenter power use is forecast to rise **165 percent globally** by 2030, from 1–2 percent of global electricity in 2023 to between 3 and 4 percent by the decade's end.

Why modernization is worth the effort

Investments in AI infrastructure highlight the need for modern systems to support scalable AI, but many companies aren't there yet.

Legacy systems lower annual revenue by **20%-30%** and consume 60%-80% of IT budgets. Modernization addresses this by transforming existing applications and infrastructure into higher-value cloud native architectures, allowing teams to deliver value faster. A joint **AWS-Coursera study found 94%** of senior tech executives see modernization as a top priority. Yet, 71% are only at the start of their journey.

Historically, modernization took 2-3 years and progressed through:



How AI accelerates modernization

AI in industries requires modernization. What about using AI for modernization itself? SoftServe's AIDEEQ transforms how you build, migrate, and modernize. Think of it as the difference between a hammer and a nail gun. It runs with existing SDLC processes, which improves the developer's experience by allowing natural interaction with systems and platforms.

Likewise, AI tools like AWS Transform, Amazon Q Developer, and Kiro automate steps, for example, write tests, generate documentation, and convert code for containerization. We're seeing an average productivity increase of 20% for developers, including a cloud-based software solutions provider for medical practices, analyzing 1.66 million lines of code in three weeks without SMEs.

How does AI impact productivity?

Genentech saved nearly five years of manual effort by deploying generative agents on AWS.

A fleet management company cut workload by 30%, costs by 35%, and security incidents by 40%.

ServiceTrade's Gen AI pilot on Amazon Bedrock increased technician productivity — leading to a better customer experience.

Section 2: Build a Scalable AI Strategy

We've shown how modern cloud and AI architecture help you get the most out of your investments and lower costs. Here's how you can bring this to life at your company.

Cloud-native AI (CNAI) is the next step. CNAI is a newer concept, and not everyone knows what it means. At its core, CNAI is about building and running AI in a way that takes full advantage of the cloud.

For leaders, the value lies in what it delivers. With CNAI, companies deploy faster, reduce costs, and scale securely, without building from scratch. Here's what CNAI makes possible, and what happens without it.

WITHOUT CNAI	WITH CNAI
~~One-off pilots that don't scale~~	Repeatable AI workflows
~~Fragmented, risky implementation~~	Secure by design (governance, compliance)
~~Siloed, hard to integrate models~~	Scalable across teams and use cases
~~Long development cycles~~	Faster time-to-value
~~High maintenance and rework~~	Lower TCO
~~Built on legacy and ad hoc infrastructure~~	Build on cloud-native services

What are CNAI principles?

1. **Serverless-first architecture:** Scales automatically based on demand, keeping costs low with pay-per-use models.
2. **Built-in governance:** Embedded security, compliance, and monitoring capabilities from the ground up.
3. **Microservices integration:** Building with microservices, containers, and serverless architectures helps reduce your attack surface. Each part is smaller, isolated, and easier to secure.
4. **API-driven connectivity:** The new API-based deployment method makes it easier to integrate with other agents and tools that use the Model Context Protocol (MCP) and Agent2Agent protocol (A2A).

We see CNAI as the blueprint for scalable AI. Our customers use services like Amazon Bedrock AgentCore, Amazon EKS, and SageMaker to build repeatable, secure AI workflows across teams and geographies. It provides necessary computing power, has shared security responsibilities, and saves money through a pay-as-you-go model. It also ensures resources are available on demand and avoids lengthy planning.

Success across industries

Financial services, healthcare, and retail lead CNAI adoption through focused use cases, strong data strategies, and executive commitment. Consider **Georgia-Pacific**, which built a cloud-native data lake on Amazon S3, incorporating real-time data pipelines and governed access.

Manufacturing is also seeing unexpected value. ENGIE Digital cut compute costs by 90% with CNAI-powered predictive maintenance and AWS solutions. **Learn how.**

Predicted equipment failures 60–90 days in advance

Reduced paper trail by 40% on one line

Increased profits by millions of dollars

Platform engineering turns CNAI from idea into a solution

CNAI delivers, but only if it's built on the right foundation. Platform engineering is that foundation. Think of it as the internal highway system that lets developers move fast and safely.

2024

Gartner named platform engineering a top trend.

2026

80% of large software engineering companies will have platform engineering teams (up from 45% in 2022).

2027

70% of companies of those teams will include Gen AI capabilities.

Most companies run in a hybrid environment and have **talent gaps**. Platform engineering (with **reusable components**, **self-service tools**, and built-in governance) bridges these gaps.

CNAI supports AI workloads. AI workloads are heavier and structurally different. They require dynamic scaling, experimentation tools, and governance frameworks that traditional workloads don't require. Platform engineering supports those needs.

Platform engineering without the pain

Switching to platform engineering takes time, talent, and coordination across teams. **SoftServe Adaptive Modernization Platform (SAMP)** is one example of a platform that helps you switch faster. Built-in dashboards, reusable components, and self-service tools give visibility and control. Use these results to justify investment and build internal momentum.

Under

20

minutes to deploy

Up to

30%

lower cloud costs

20%-30%

boost in developer productivity with Gen AI

4-6

months of acceleration with AI agents and automation

A 6-month success story with SAMP

SonicWall partnered with SoftServe and AWS to create a shared services platform and accelerate its modernization timeline by six months.

SAMP offers automation, ease of use, and customizability. Fixed cost, fixed duration, fixed deliverables with success stories to back it up. And, it's built for AI use cases with trusted tech partners. Built on AWS-native services and **NVIDIA Blueprints**, SAMP delivers industry-specific solutions that increase productivity and accelerate AI adoption.

Section 3: Tips for Better ROI in AI

Many companies stall in the pilot phase, believing AI is too complex to deliver business value. With the right approach and partners, AI drives measurable returns — faster innovation, improved efficiency, and stronger competitive advantage. We've helped clients unlock ROI with four strategies:

Why it's affordable

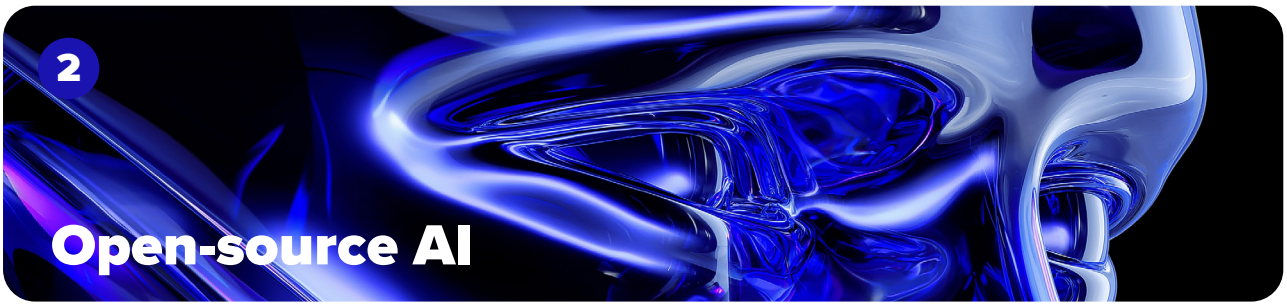
Reduces operational overhead and accelerates time-to-value.

Business value

Proves value early, minimizes risk, and supports scalable AI adoption.

How it works

Services like Amazon Bedrock allow teams to quickly evaluate foundation models without managing infrastructure. This supports fast experimentation and cost-optimized model selection.



2 Open-source AI

Why it's affordable

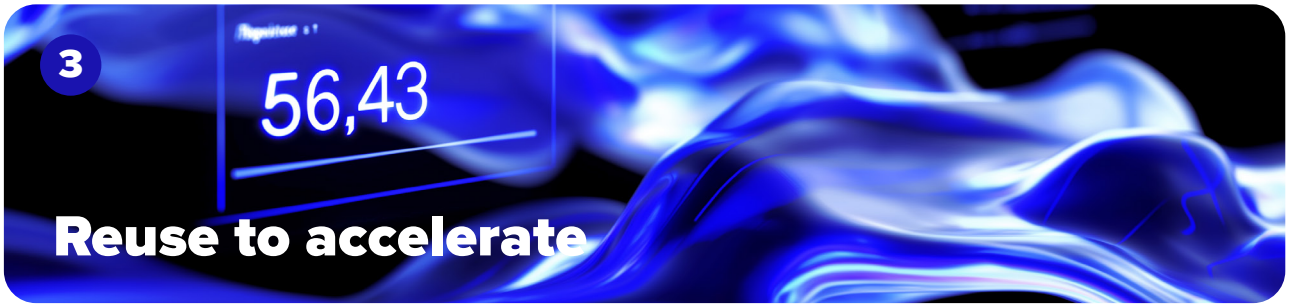
No licensing fees, broad community support, and flexibility to customize.

Business value

Accelerates innovation and lowers the threshold for contributing to AI initiatives. It also reduces development costs while maintaining enterprise-grade performance and security.

How it works

AWS provides frameworks like Hugging Face and PyTorch (e.g., security controls, compliance certifications, and governance tools) as well as open-source models such as DeepSeek-R1 through Amazon Bedrock or on Amazon Elastic Kubernetes Service (Amazon EKS).



3 Reuse to accelerate

Why it's affordable

Cuts time-to-value and avoids duplication of effort.

Business value

Minimizes rework, speeds up deployment, and scales AI faster.

How it works

SAMP, for example, provides reusable blueprints, self-service platforms, and modular components tailored to your environment and needs.

4

NVIDIA-accelerated solutions and AWS purpose-built silicon

Why it's affordable

AWS offers superior price-performance alternatives with custom silicon designed for AI workloads. AWS Trainium-based Amazon EC2 Trn1 instances reduce training costs by up to 50% compared to GPU-based instances, while AWS Inferentia-powered EC2 Inf1 instances achieve up to 2.3x higher throughput and reduce cost per inference by up to 70% versus comparable Amazon EC2 instances.

Business value

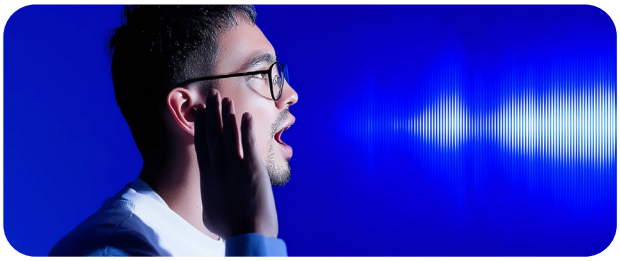
High-performance AI (e.g., speech, multimodal, content generation) with lower TCO. Gives teams flexibility to choose the most cost-effective compute option for their use case. AWS promotes a vertically integrated AI infrastructure built for scale, performance, and cost-efficiency. Companies can optimize their AI workloads based on specific performance and budget requirements.

How it works

As an elite partner, SoftServe integrates NVIDIA frameworks (e.g., NeMo, Riva, TensorRT) with AWS infrastructure to deliver scalable, high-performance AI solutions. AWS also maintains deep collaboration with NVIDIA as one of its largest cloud partners, offering GPU-powered instances through EC2 and Amazon SageMaker.

The second-generation Inferentia2 provides up to 4x higher throughput and up to 10x lower latency, making it perfect for inference tasks using LLMs or diffusion models. This provides teams with flexibility to choose between high-performance NVIDIA GPUs for the most demanding workloads or cost-optimized AWS silicon for production-scale deployments.

Use Cases



Speech Recognition Platform

Designed for children's speech, achieving a word error rate 3x lower than other ASR systems.

Digital Concierge

Avatar-based Gen AI system for real-time customer support, reducing **operational costs by up to 40%**.

Multimodal RAG System

Integrates text, images, and diagrams for context-aware responses across industries.

Content Creator

Automates visual content generation using brand-specific data and A/B testing.

Section 4: Measure Success

To measure AI success, track KPIs that reflect business impact, not just infrastructure costs. Here's what leading companies measure:

Total cost of ownership (TCO): GE Oil & Gas reduced TCO by over 50%.

Business agility: Unilever launched new products 75% faster with AWS solutions.

Speed: A networking and IT infrastructure company achieved cloud-agnostic modernization in under three months with SAMP.

Innovation velocity: HubSpot scaled its image generation capabilities by 150%.

Operational resilience: Expedia Group ensured critical workloads ran across multiple AZs and regions for robust disaster recovery with AWS solutions.

Staff productivity: A fleet management company reduced workload by 30% with SoftServe SAMP, while **Sage Software** saved over 500 hours per year in server configuration time with AWS solutions. Developer improvements translate into concrete, measurable value. Amazon saw a **15.9% year-over-year reduction** in software delivery costs.

For a closer look at the cost structures, scaling dynamics, and business models emerging around agentic AI, see our companion report

Economics of Agentic AI

Start with the business goal, not the experiment

We touched on AI purgatory at the start, and it's applicable here. AI projects, like hackathons orproofs-of-concept, fail because they don't have a clear goal.

Tip

To build a strong business case, you first need to define your desired outcome. Then, use AI tools (like AWS **Cloud Value Framework**) to assess the total cost and ROI of a repeatable solution that achieves that outcome.

Once your goal is clear, shift from isolated projects to measurable progress. Leading companies use KPI-based frameworks to track impact. Choose metrics that reflect your business priorities: revenue growth, customer retention, operational efficiency, or a mix of all three. Decide first, then measure constantly. With techniques like A/B testing, real-time analytics, and feedback loops, teams can validate what works and scale it faster.

For example, **Warner Bros. Discovery** used an AI promotion engine for brands like TBS, TNT, and Adult Swim, leading to a 14% increase in engagement and a 12% lift in cross-brand interaction.

And, if you're just getting started, SoftServe's "**10 Essential Tactics To Improve Your AI Strategy**" recommends you pick low-risk pilots, like automating data entry, before you move to complex scenarios such as predictive analytics.

Agentic AI is moving fast

Scaling AI over the next 5-10 years requires a flexible, automated cloud foundation, especially to support agentic AI.

Agentic AI systems independently understand, reason, and act over time to achieve goals. Unlike traditional applications, these systems make dynamic decisions, adapt to information, and typically require tools, APIs, and contextual memory. Modernize older applications to create modular service agents that you can use within broader workflows.

Services like Amazon Bedrock AgentCore Gateway accelerate the development of agentic systems. They connect tools and APIs and work with serverless platforms like AWS Lambda. This allows for dynamic execution and saves money. To prepare for this future, companies should focus on:

Serverless infrastructure

Scale workloads as needed for fast experimentation and cost control.

Automated MLOps

Simplify the AI life cycle (from data to deployment) to reduce time-to-value.

Responsible AI frameworks

Build in governance, fairness, and compliance to ensure trust and regulation.

Edge AI capabilities

Get real-time decisions in factories, vehicles, and retail. Lower latency, higher impact.

Make a plan

We've covered the full landscape, from modernization strategies to platform engineering and measurable outcomes. Successful AI isn't luck. It's built on planning, implementation, and continuous improvement. Develop a plan to bring AI into your operations, using cloud technologies to achieve this. Modernize what's important and migrate what makes sense.

Modernize your cloud for AI success: 5 Takeaways

1. Start with a business goal, not an experiment
2. Modernize your data foundation
3. Adopt CNAI principles
4. Empower teams with platform engineering
5. Measure what matters

SoftServe and AWS make modernization faster

For over a decade, AWS and SoftServe have partnered to help enterprises modernize faster and securely scale AI. We combine AWS' cloud innovation with SoftServe's engineering expertise to deliver tangible results.

Our collaboration has produced more than 100 joint proofs-of-concept, launched early-access programs in Generative and agentic AI, and accelerated modernization through SAMP. Clients have seen up to 30% cost savings and 5x faster delivery.

Through our **Strategic Collaboration Agreement (SCA),** we continue to co-create, experiment, and deliver measurable outcomes, from financial services to healthcare.

The future of AI is built on cloud-native foundations. With AWS and SoftServe, you're not just adopting technology; you're building a foundation for intelligent transformation.

Let's build your five-year AI plan together.

Contact Us

About SoftServe

SoftServe is a premier IT consulting and digital services provider. We expand the horizon of new technologies to solve today's complex business challenges and achieve meaningful outcomes for our clients.

Visit SoftServe's **website, blog, LinkedIn, Facebook,** and **X (Twitter)** pages for more.

About AWS

Amazon Web Services is the world's most comprehensive and broadly adopted cloud, enabling customers to build anything they can imagine. We offer the greatest choice of innovative cloud capabilities and expertise, on the most extensive global infrastructure with industry-leading security, reliability, and performance.

info@softserveinc.com
www.softserveinc.com

softserve | aws