CYBAGE
Delivering Value. Scientifically.

aws

Open AI Migration eBook

Agenda

Part 1: The Strategic Landscape

1. Introduction: Why AI Workloads Choose AWS
2. Why Customers Prefer AWS for Cloud AI Solutions
3. The Cybage Advantage
4. Gen AI Migration Journey: A Strategic Overview

Part 2: AWS AI Services in Action

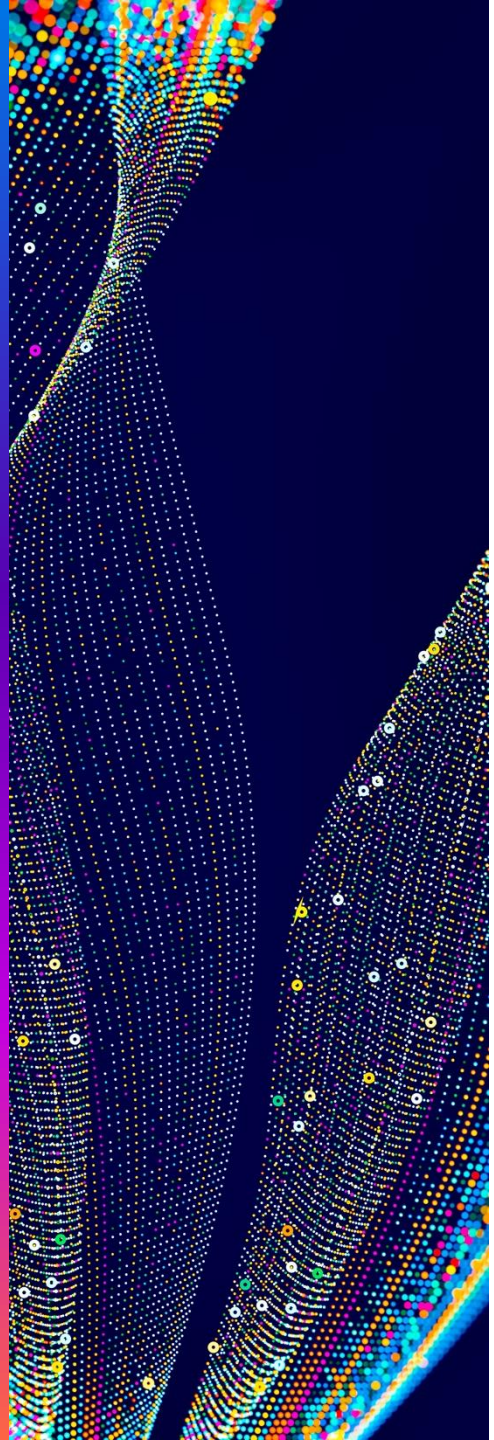
5. Amazon Bedrock: Capabilities & Key Benefits
6. Amazon Nova: Features & Key Benefits
7. Amazon Nova Foundation Models Explained
8. Migrating from OpenAI to Amazon Nova: Why and How
9. Amazon SageMaker for AI: Benefits for Scalable ML
10. Cybersecurity Reinvented with AWS AI

CYBAGE
Delivering Value. Scientifically.

aws

Part-1

The Strategic Landscape



Companies running on AI workloads frequently ask us:

- Is this solution production-grade and stable?
- Can we optimize for latency and responsiveness at scale?
- What's the cost-to-serve between different models?
- Can we monitor, troubleshoot, and iterate with best practices?
- How do we ensure trust, governance, and compliance?
- How do we continuously measure and improve quality post-launch?
- What migration incentives and programs are available?

90%
of gen AI prototypes never make it to production

74%
of gen AI projects hit performance or reliability issues in production

74%
of companies report lack of enterprise-grade security controls[1]

>6
seconds of latency leads to drop-offs in customer adoption

1 Bain & Company Enterprise GenAI Adoption Review

Why customers choose AWS

Cost optimization and efficiency

- Best performance and cost for gen AI services with flexible pricing models
- Consumption of services (including models on Amazon Bedrock) draws from EDPs

Uncompromised security and governance

- Full enterprise compliance (SOC 2, HIPAA, FedRAMP)
- Built-in guardrails and security/compliance controls
- Full control of your data; nothing stored or ingested by LLM model providers

Unified model access and tooling

- Single API access to leading models (Anthropic, Llama, Cohere, etc.)
- Consistent benchmarking and evaluation tools for your use cases
- Bedrock Studio for streamlined development

Managed infrastructure and integration

- Native AWS service integration; integrates with your existing security/networking and compliance on AWS
- Pre-built connectors for enterprise systems

Easy to start, deploy, and maintain

- Hosting data, apps, and gen AI services within the same cloud simplifies system maintainability and ability to innovate

Customer success is the highest priority

- AWS Builder teams and APN Partners are obsessed with helping customers succeed with gen AI

Why Cybage

A recognized leader in Gen AI consulting solutions / GenAI solution engineering and strategic AWS partner

- **GenAI Migration & Modernization**: Expert in migrating enterprise GenAI workloads to AWS-native stacks
- **Faster, Safer Builds**: Accelerated Bedrock builds using RAG and SmartPal frameworks
- **Governance-first AI Services Provider**: We embed observability, RBAC, and compliance at scale

AWS Advanced Tier Services Partner

with 100+ successful engagements, Cybage accelerates enterprise digital transformation with cloud-native solutions.

1500+
AWS Skilled Personnel

75+
AWS Accredited / Certified Individuals

06
AWS competencies

100+
customers served on AWS

The Strategic Collaboration Agreement with Amazon Web Services (AWS)—a move designed to fast-track enterprise GenAI adoption with security, speed, and scale.

News Update:
Cybage Signs Strategic Collaboration Agreement with AWS to Drive Scalable, Responsible Gen AI

Gen AI migration journey

1 Discovery & qualification

Objective:
Customer pain points and migration triggers identified

2 Assessment & planning

Objective:
Develop migration strategy

- Technical architecture documented
- Performance requirements captured
- Migration approach defined

3 Evaluation & proof

Objective:
Validate AWS solution meets requirements

- Model evaluation
- Performance benchmarks conducted
- Cost projections created
- Technical feasibility confirmed

4 Migration execution

Objective:
Execute migration to AWS

- Migration plan finalized Technical implementation completed
- Testing & validation performed
- Commercial and legal guidance

5 Continuous collaboration

Objective:
Ensure customer success

- Production metrics validated
- Continuous support provided
- Additional use cases identified

Migrations are classified into three categories based on patterns:

Simple API endpoint switching

Timeframe: 1–4 weeks

Transitions from externally hosted endpoints to Bedrock with minimal technical effort.

Advanced workload migration

Timeframe: 1–3 months

Implementations involving custom components, including Custom Model Import (CMI), fine-tuning, and deployment of custom models on Bedrock.

Full-stack application migration

Timeframe: 4–6 months

Complete application transitions including data dependencies, Retrieval Augmented Generation (RAG) components, agents, and other application elements requiring significant refactoring.

CYBAGE
Delivering Value. Scientifically.

aws

Part-2

AWS AI Services in Action

Amazon Bedrock

Managed service to build and scale gen AI applications with powerful tools and a broad range of foundation models

Amazon Bedrock: Key benefits

- Choice of leading FMs through a single API
- Secure model customization within the development environment
- Ability to create managed agents that execute multistep tasks
- Security, privacy, and data governance
- Serverless — no infrastructure to manage

Simplify development with model choice, customization and integration while maintaining privacy and security.

Amazon Nova

State-of-the-art foundation models
that deliver frontier intelligence and industry
leading price performance

Amazon Nova: Key benefits

Cost savings

With token costs up to 4x lower than competitors, businesses can scale applications more economically.

Enhanced response performance

Faster response times (up to 190 tokens per second) make real-time applications more viable.

Expanded capabilities

A larger context window and multimodal support unlock new applications, from detailed document analysis to integrated visual content.

Key features:

- Integrates with Amazon Bedrock
- Fully customizable
- State-of-the-art video understanding capability
- Powers agentic workloads
- Responsible AI

Amazon Nova foundation models

UNDERSTANDING MODELS

Amazon Nova Micro

Our text only model that delivers the lowest latency responses at very low cost

GENERALLY AVAILABLE

Amazon Nova Lite

Our lowest cost multimodal model that is lightning fast for lightweight tasks

GENERALLY AVAILABLE

Amazon Nova Pro

Our highly capable multimodal model with best combination of accuracy, speed, and cost for a wide range of tasks

GENERALLY AVAILABLE

Amazon Nova Premier

Our most capable multimodal model for complex reasoning tasks and for use as the best teacher for distilling custom models

COMING SOON

Lower Cost & Latency ⟷ *Increasing Intelligence*

CREATIVE CONTENT GENERATION MODELS

Amazon Nova Canvas

State-of-the-art image generation model

GENERALLY AVAILABLE

Amazon Nova Reel

State-of-the-art video generation model

GENERALLY AVAILABLE

Why migrate from OpenAI to Amazon Nova?

Model	Input Token Cost (per Million Tokens)	Input Token Cost (per Million Tokens)	Context Window	Output Speed (Tokens per Second)	Latency (Seconds per first token)
GPT-4o	~$2.50	~$10.00	Up to 128K tokens	~63	~0.49
GPT-4° Mini	~$0.15	~$0.60	Up to 128K tokens	~90	~0.43
Nova Micro	~$0.035	~$0.14	Up to 128K tokens	~195	~0.29
Nova Lite	~$0.06	~$0.24	Up to 300K tokens	~146	~0.29
Nova Pro	~$0.80	~$3.20	Up to 300K tokens	~90	~0.34

Key Takeaways:

- AWS Nova Pro costs **68% less** than GPT-4o for processing 20M tokens.
- AWS Nova Pro has a **larger context window** (300,000 vs. 128,000 tokens).
- AWS Nova Pro processes tokens **twice as fast** (157 vs. 77.4 tokens/sec).

68% less!

Check out this blog for details

Amazon SageMaker

Managed, end-to-end service to build, train, and deploy ML models at scale, including foundation models

Amazon SageMaker for AI: Key benefits

Diversification of FMs

Access over 250+ of the latest and publicly available FMs, architectures, and modalities

Strong control and flexibility

Control over FM development and hosting

Customization at scale

Build FMs from scratch and advanced FM fine-tuning techniques

Optimize cost and performance

Lower costs for re-training and hosting FMs

End-to-end FM lifecycle management

FM training, customization, deployment, and inference

Implement FMOps and governance

The most secure and reliable fully managed infrastructure with scalability and high-performance

Case Study

> We needed to rapidly develop and deploy a secure, scalable chatbot for our managed service providers. Cybage helped us pivot from open-source solutions to AWS-native services, enhancing our security posture and reducing operational complexity. With their guidance and partnership with AWS, we successfully rearchitected and hardened our existing prototype on a short timeline before GA. This has been instrumental in accelerating our AI transformation journey.
>
> — Cloud Engineering, SRE & FinOps Leader

GenAI at Scale: A Cybersecurity Leader's Move to Bedrock-Powered Automation

How a Cybersecurity Leader Streamlined Firewall Ops with Multi-Agent GenAI on Amazon Bedrock

About the Client

A multi-billion-dollar cybersecurity powerhouse protecting enterprises worldwide with next-generation firewalls and cloud-native security solutions. Their infrastructure processes millions of security events daily, making real-time decision-making critical to business success.

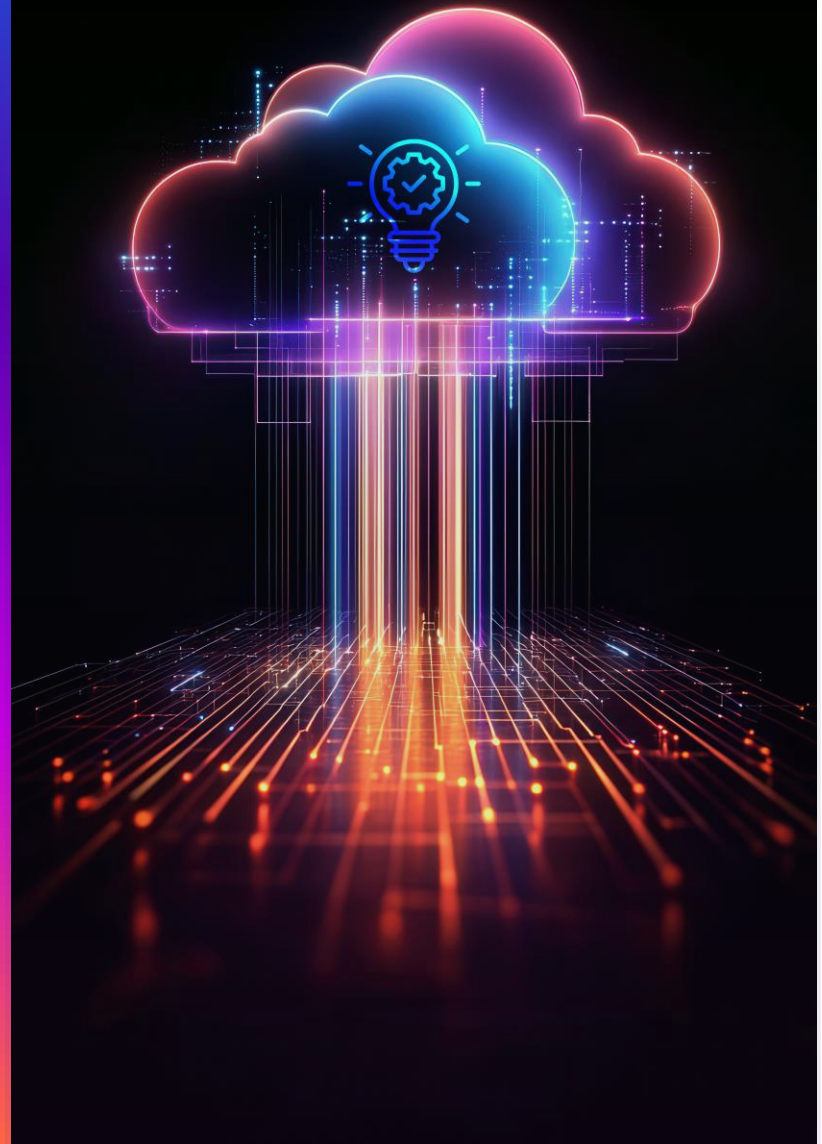
The Mission

From Clicks to Commands

Firewall operations were slow and manual. Analysts had to click through multiple dashboards just to check statuses, pull reports, or run basic actions. Each task took time, adding up to 30 hours a week per person.

The mission was clear: replace clunky workflows with fast, natural language interactions. Ask a question. Get an answer. Take action. Enable instant troubleshooting, proactive compliance monitoring, and real-time insights while maintaining enterprise-grade security standards. All within a secure, AWS-native environment.

Case Study



From Multi-cloud Complexity to AWS-Native Control

Business Challenge

Multi-Cloud Operational Complexity

While the client's initial Azure OpenAI proof-of-concept demonstrated strong functionality, it exposed significant operational and technical challenges:

- Security Risks from cross-cloud data flows
- 40% higher OpEx to support dual-cloud operations
- Integration overhead and fragmented tech stack

The Solution

AWS-Native AI Architecture

Cybage engineered a modular, production-grade multi-agent architecture using Amazon Bedrock and Claude 3.5, layered with secure, real-time backend integrations.

Key capabilities:

- Natural Language Queries for real-time status, updates, expiries
- Automated Troubleshooting for faster issue resolution
- Compliance Checks to ensure upgrade and policy adherence
- Real-time Data Insights with secure backend integrations

The Strategic Catalyst

The launch of function-calling capabilities by Claude on Amazon Bedrock presented an opportunity to consolidate platforms and go all-in on AWS-native GenAI.

Case Study

The Payoff: Scalable Automation, Stronger Security & Reduced Costs

The Migration

Model Migration

Azure Open AI → Claude on Amazon Bedrock

Library Shift

LangChain → Boto3 for tighter AWS control

Outcome

Simplified architecture, unified governance, real-time automation

Under the Hood

Bedrock Knowledge Base	OpenSearch Serverless	Anthropic Claude 3.5	Systems Manager	AWS Bedrock
Redis	PostgreSQL	Boto3 SDK	Langfuse	EKS	S3

The Results

Enterprise Impact at Scale

- **Improved Security Posture**: Eliminated cross-cloud data flows
- **40% Cost Reduction**: Single-cloud architecture efficiency
- **Performance at Scale**: Real-time processing with a scalable multi-agent framework
- **3x Faster Development**: Simplified AWS-native integrations
- **70% Analyst Time Reduction**: ~30 hours/week saved

CYBAGE
Delivering Value. Scientifically.

aws

What This Means For You

If GenAI is siloed or slowing you down, it's time to move to a scalable, secure AWS-native architecture.

Let's transform one high-friction workflow for you - securely and fast.

Book your 30-minute GenAI productivity assessment with Cybage now.

business@cybage.com | *Explore more*