

Amazon SageMaker

The center for data, analytics & AI

Anton Kartashov

Solutions Architect

Amazon Web Services | WWPS CEE

E: antkar@amazon.com

P: +41 76 328 75 44

[linkedin.com/in/anton-kartashov](https://www.linkedin.com/in/anton-kartashov)



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Barriers to adoption
& customer needs

SageMaker AI overview

Core capabilities

Customers stories

Getting started



Hundreds of thousands of customers



Times are changing

Technology



Changing faster than
ever

Work



Work and interact
with technology

Organizations



Adopt and adapt

Organizations understand that

Analytics & AI

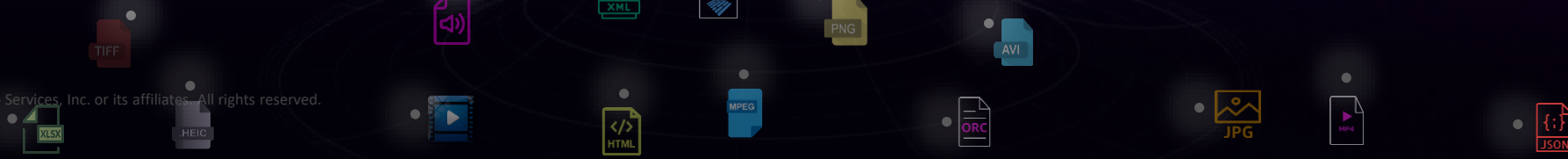
tailored to their business

fuels

Their **Data**



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Organizations are struggling

74% of businesses fail to turn data into insights



82%

of organizations have appointed a Chief Data Officer (CDO)



94%

of organizations increased data investments in 2023



20%

revenue growth seen by data-insights driven businesses

Challenges in an evolving landscape

High effort AI/ML

AI and machine learning faces cost, speed, and efficiency barriers

Evolving personas

Converging personas and jobs to be done across AI/ML and analytics

Evolving tech environment

Data silos

Data silos that are not interoperable with the tools

Fragmented governance

Trust, confidence, and compliance in data management

Evolving data environment

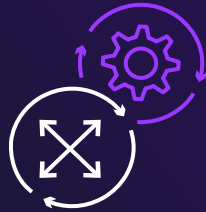
Evolving technology is leading to complex tool-set

Organizations are struggling with data sprawl

What customers are asking...



**How do I choose
between an
existing model and
building one?**



**How do I
optimize
training and
inference
cost?**



**How to improve
accuracy
while scaling
model size?**



**How can I deploy
ML and Foundation
Models at
scale?**

Organizations appreciate

Comprehensive set of purpose-built services

Optimized for **performance and cost**

Experience



Act



Govern



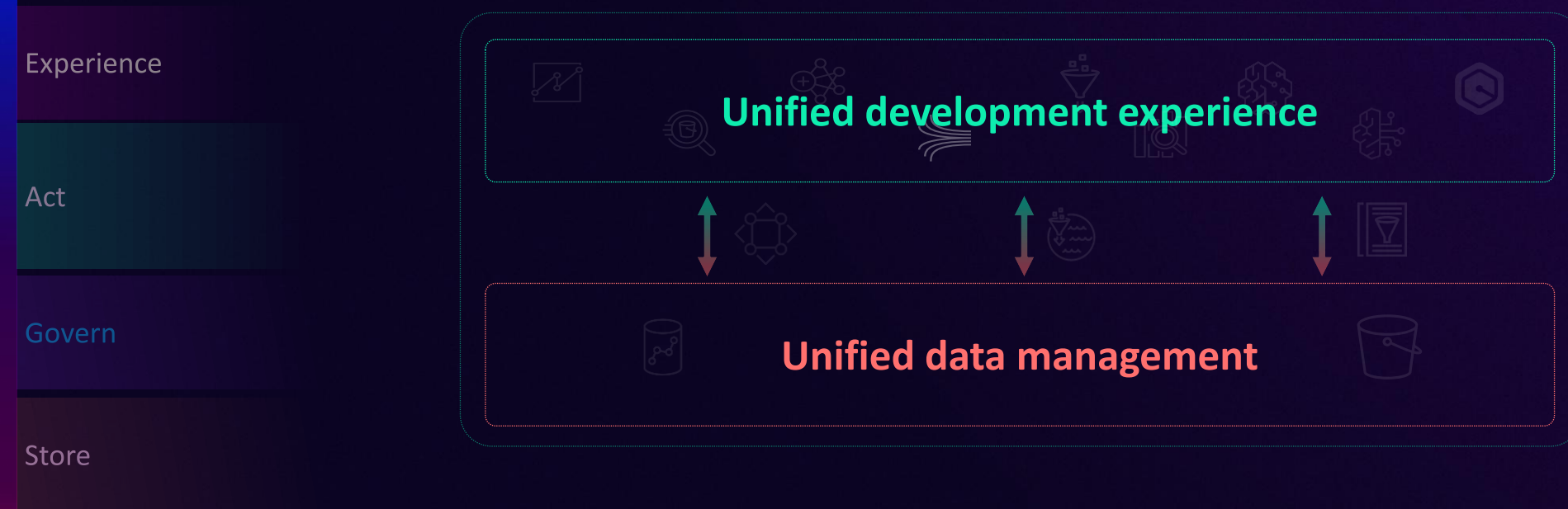
Store



Organizations want

Leverage same **rich set** of services

through **unified experiences** across them



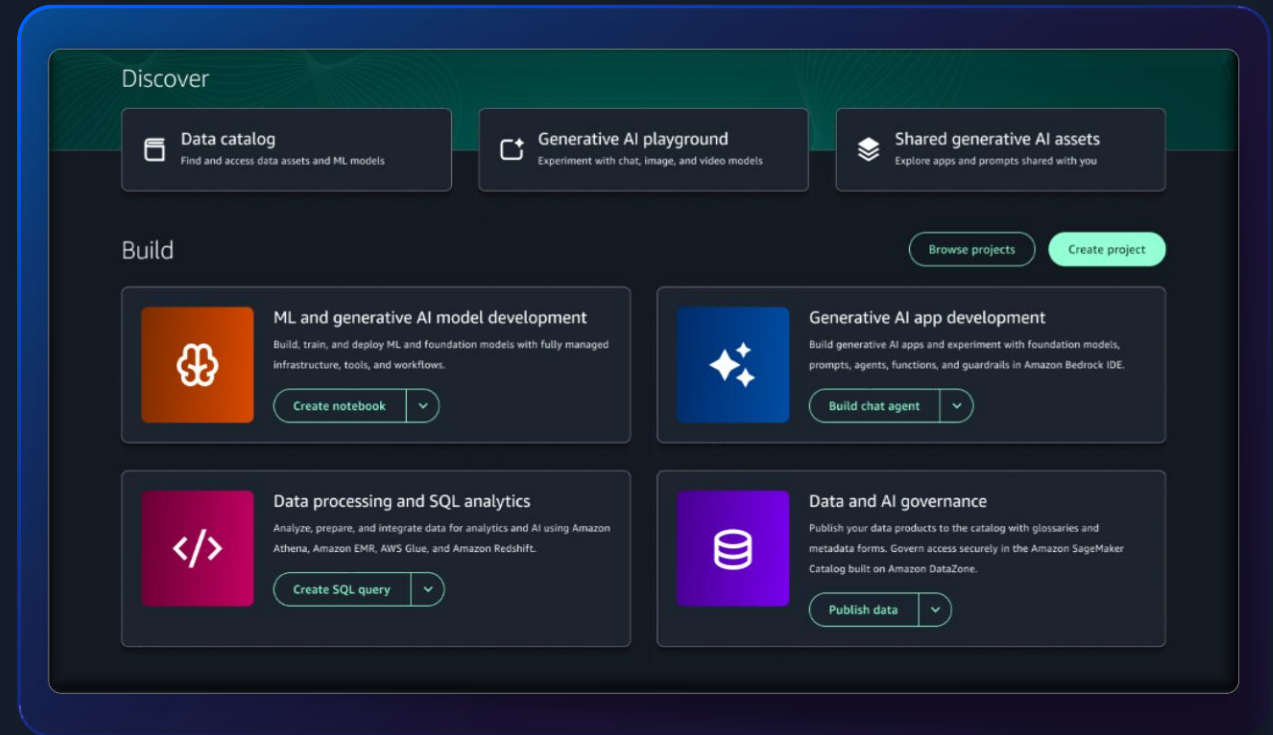
Amazon SageMaker Unified Studio

ACCESS ALL YOUR DATA AND TOOLS FOR ANALYTICS AND AI IN A SINGLE ENVIRONMENT

Integrated experience for data preparation, model building, and generative AI application development

Unifying your tools such as notebooks and query editors across services

Seamless integration with AWS data processing, analytics and ML services like EMR, Glue, Athena, Redshift, and Bedrock

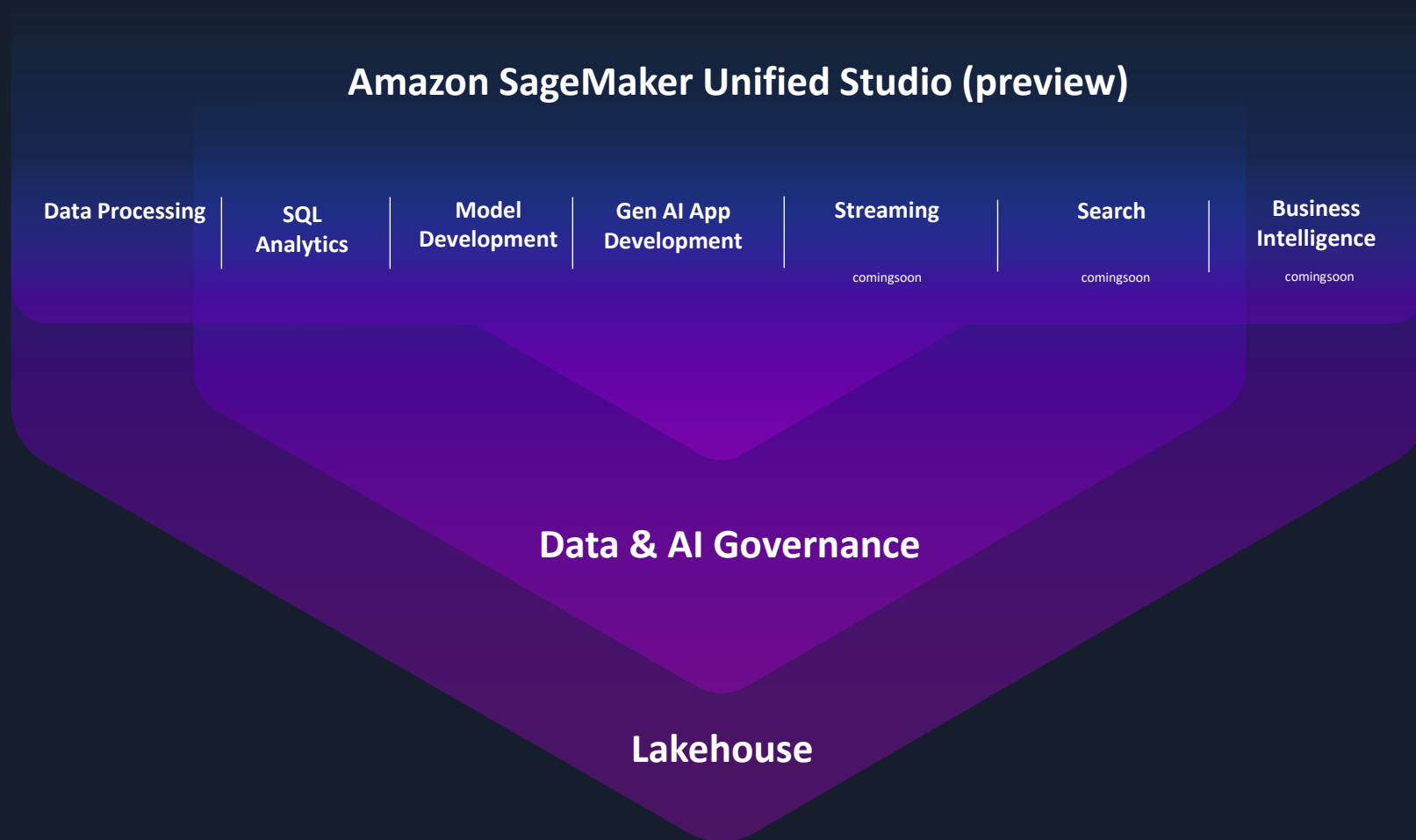


Amazon SageMaker Unified Studio (Public Preview)



Amazon SageMaker Unified Studio

ACCESS ALL YOUR DATA AND TOOLS FOR ANALYTICS AND AI IN A SINGLE ENVIRONMENT





Amazon SageMaker AI

Build, train, and deploy ML models at scale, including FMs

Build FMs from scratch

Customize FMs

Access the latest and publicly available FMs

Manage and deploy models for inference

Implement FMOps and governance



Amazon SageMaker AI

Build, train, and deploy ML models at scale, including FMs



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Build FM's from scratch

Create your own ML models, including FMs, with integrated purpose-built tools and high-performance, cost-effective infrastructure



Customize foundation models

Access and evaluate 250+ FMs that can be customized easily for your use case



Implement MLOps & governance

Create reliable and repeatable workflows incorporating MLOps practices with purpose-built tooling



Improve ML governance

Enhance model governance and compliance with built-in governance tools



Manage and deploy models for inference

Easiest way to deploy AI & ML models including foundation models (FMs) to make inference requests at the best price performance for any use case

SageMaker AI supports ML, Deep Learning and GenAI

AMAZON SAGEMAKER AI

Machine Learning (Tabular Inputs)

Predictive maintenance
Financial risk prediction
Demand forecasting
Fraud detection
Churn prediction
Personalized recommendations

Deep Learning (Unstructured Inputs)

Computer vision
Meta data enrichment
Sentiment analysis
Topic modelling
Intelligent data processing
Autonomous driving

Gen AI (Unstructured outputs)

Summarization
Information extraction
Visual content generation
Code generation
Audio/music generation
Synthetic data generation

Classic ML and GenAI with Amazon SageMaker AI

Classic ML



Data Prep

Prepare data assets and pipelines,
Manage data quality and bias



Build

Experiment and automate the
execution of build pipelines



Train

Train ML models at scale
with automation



Deploy

Automate the execution of
deploy pipelines into production

GenAI



Data Prep



Pre-train or Select

Select from FM hub or
BYOFM, or Pre-train FM



Evaluate

Automated or human evaluation



Fine-tune

Fine-tune and/or optimize
models for deployment



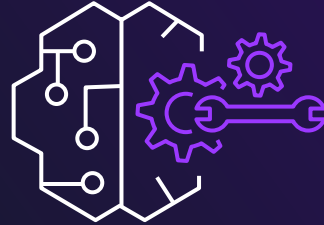
Deploy

AUTOMATE WITH AI OPS AND GOVERNANCE

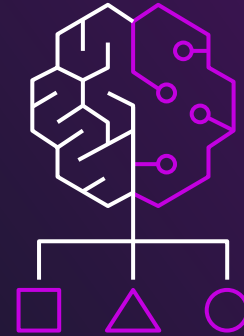
Amazon SageMaker AI simplifies



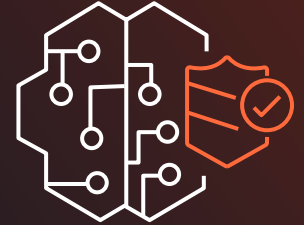
**Model
Building**



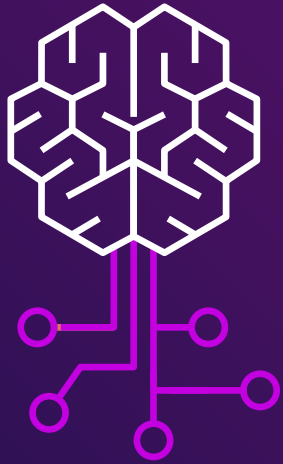
**Model Training
& Fine-Tuning**



**Model
Deployment**



**Scalable MLOps
and governance**



Model Building

Single, fully managed IDE for notebooks, code, and data



Accelerate and scale data prep for AI and ML

Use the most comprehensive
of tools for both structured
unstructured datasets

set
and



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Access data

Easily access and query data from a wide variety of data sources



Cleanse, label, enrich

Create high quality labeled datasets for training models using your tool of choice or through human feedback



Analyze and visualize

Explore data through purpose-built analysis and visualization tools, or visualize geospatial data on interactive 3D accelerated maps



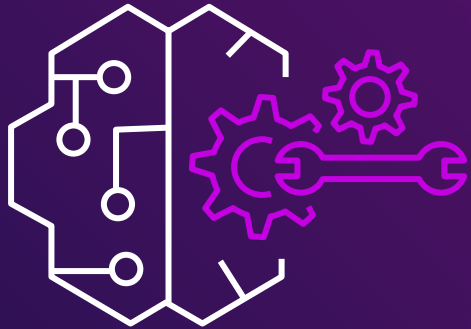
Scale

Efficiently process large amounts of data while reducing cost



Store and share

Securely store, manage, and share features to be used the ML lifecycle



Model Training & Fine-Tuning

Fast and cost-efficient ML and Gen AI model training



Amazon SageMaker JumpStart

Model hub with foundation models, built-in algorithms, and prebuilt AIML solutions that you can deploy with just a few clicks



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



ML hub with foundation models

Access hundreds of foundation models including top open-weight and proprietary models that can be fine tuned easily for your use case



Ease of use

Easily use pre-trained models on SageMaker AI instances like Inf2 and optimized hosting configurations through presets



Evaluate and customize

Evaluate, fine-tune, and optimize deployment using a few clicks



Data security and access control

Keep inference and training data private and curate who can access and use models within your organization



Share and collaborate within your organization

Share models and notebooks with others within your organization, and allow them to train with their own data or deploy as-is for inferencing

Amazon SageMaker JumpStart






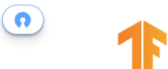






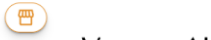



MODEL HUB WITH FOUNDATION MODELS, BUILT-IN ALGORITHMS, AND PREBUILT AIML SOLUTIONS

Over 250+ publicly available foundation models

Hundreds of built-in algorithms with pretrained models from popular model hubs

Fully customizable solutions for common use cases with reference architectures

Share AI and ML models and notebooks across your organization

 HuggingFace Explore hundreds of popular and trending models from HuggingFace. View 368 models >	 Meta Explore popular and trending models from Meta including Llama, Code Llama, and more. View 44 models >	 AI21 Explore popular and trending models from AI21 Labs including Jurassic and more. View 6 models >	 Stability AI Explore popular and trending models from Stability.ai including Stable Diffusion and more. View 11 models >
 Cohere Explore popular and trending models from Cohere including Command, Rerank, and more. View 12 models >	 TensorFlow Explore popular and trending models from TensorFlow for computer vision and NLP tasks. View 319 models >	 PyTorch Explore popular and trending models from PyTorch for computer vision and NLP tasks. View 34 models >	 Upstage Explore popular and trending models from Upstage including Solar mini chat model and more. View 4 models >
 LightOn Explore popular and trending models from LightOn including mini-instruct models. View 2 models >	 NCSOFT Explore popular and trending models from NCSOFT including VARCO LLM models. View 3 models >	 LG CNS Explore popular and trending models from LG CNS including EXAONE Atelier and more. View 1 models >	 Jina AI Explore popular and trending models from Jina AI including Jina Embeddings model and more. View 1 models >
 Voyage AI Explore popular and trending models from Voyage AI including Voyage-2 Embedding Model and more. View 3 models >	 Nomic Explore popular and trending models from Nomic including Nomic Embed Model and more. View 2 models >	 EvolutionaryScale, PBC Explore frontier, generative language models for biology from EvolutionaryScale, including ESM3. View 1 models >	 Amazon Explore popular and trending models from AWS for computer vision, NLP, and tabular tasks. View 36 models >



SageMaker AI offers two training options

PURPOSE-BUILT INFRASTRUCTURE FOR FM TRAINING

Fully managed training jobs

Fully managed resilient infrastructure for large-scale and cost-effective training

Focus on model building rather than IT

Provide access to flexible on-demand GPU cluster with a pay as you go option



Amazon SageMaker HyperPod

Resilient and **self orchestration** infrastructure for maximum resource control

Customize and manage cluster orchestration (Slurm or EKS)

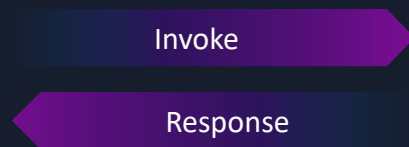
Schedule workloads to maximize cluster utilization across teams

Model Deployment

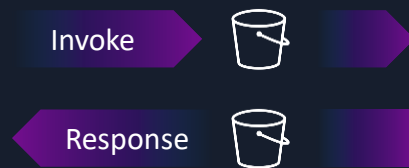
Easily deploy AI and ML models - From low latency and high throughput to long-running inference

Model deployment on Amazon SageMaker AI

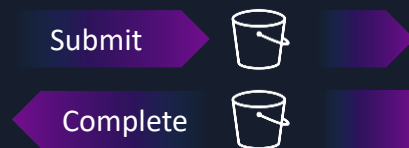
- Real-time synchronous response



- Near real-time asynchronous response



- Offline batch inference



Model



Single model
deployment



Multi-model
deployment



Multi-adapter
hosting

Container



Single container



Multi-container

TensorFlow

PyTorch

mxnet

ONNX

Keras

learn

GLUON



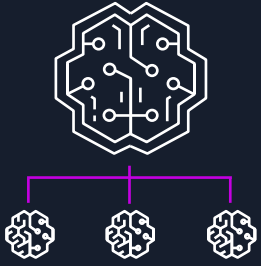
Infrastructure

Serverless

GPUs

CPUs





Deploy AI and ML models

Fully managed deployment for inference at scale



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Wide selection of infrastructure

70+ instance types with varying levels of compute and memory to meet the needs of every use case. Deliver up to 40% better inference price performance with Inf2 instances



Deploy models in production for inference for any use case

From low latency and high throughput to long-running inference



Cost-effective deployment

Reduce inference cost by at least 50% with multi-model/multi-container endpoints, serverless inference, and elastic scaling



Shadow testing & automatic deployment guardrails

Validate the performance of new ML models against production models. Minimize risk when deploying new model versions on SageMaker AI using linear, canary, or blue green traffic switching



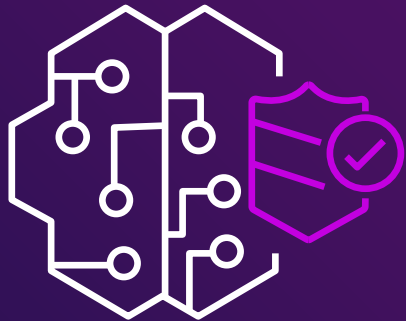
Built-in integration for MLOps

ML workflows, CI/CD, feature management, lineage tracking, and model management



Large model inference container

Achieve best price performance with the latest inference optimizations tools, model servers, and libraries packaged into a single container



Scalable MLOps

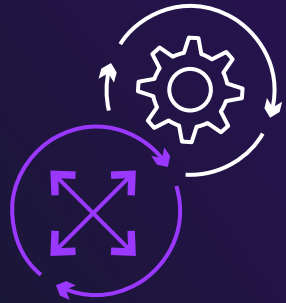
Implement reliable MLOps
workflows with built-in governance
and compliance tools

Ops challenges managing the model lifecycle

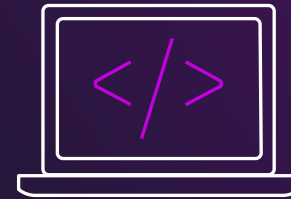
PURPOSE BUILT TOOLS FOR MLOPS AND GOVERNANCE



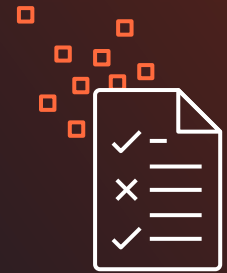
**Manual iterative
processes slow down
ML innovation**



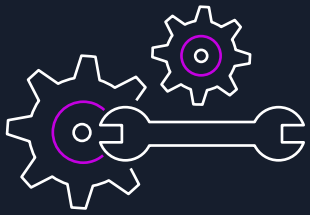
**Difficult to scale and
manage the number
of models in
production**



**CI/CD for ML
requires writing
custom code**



**Compliance
requirements
are difficult
to meet**



Amazon SageMaker MLOps

Streamline the ML lifecycle



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Automate ML workflows to scale model development



Build CI/CD pipelines for gen AI and ML to improve reliability, quality, and accelerate model deployment



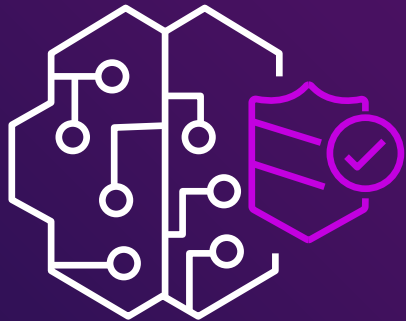
Catalog model versions, metadata, metrics, and approvals for traceability and reusability



Track lineage for traceability and compliance



Maintain accuracy of predictions after models are deployed



Built-in Governance

Simplify access control and
enhance transparency

Comprehensive data protection and privacy

PURPOSE BUILT TOOLS FOR MLOPS AND GOVERNANCE



Your data used with Amazon SageMaker AI is not used for service improvement and not shared with third-party model providers



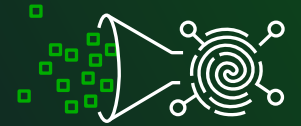
Private connectivity between Amazon SageMaker and your virtual private cloud (VPC)



Your data is encrypted in transit and at rest



Customize FMs privately, retaining control over how your data is used and encrypted



Used and encrypted Deploy FMs on single tenant endpoints provisioned for your inference use



SageMaker AI Customer Stories

NFL is using data to change the game

LEVERAGING ML TO ENSURE OPTIMAL PLAYER PERFORMANCE AND SAFETY

Challenge

NFL wanted to track every play in every NFL game and find new way to use ML to extract deeper insights to help fans, broadcasters, coaches, and teams

However, building these capabilities requires large amounts of accurately labeled training data

Solution

NFL leveraged video object tracking workflow, in addition to other computer vision labeling workflows, to develop labels for training a computer vision system that tracks all 22 players as they move on the field during plays

Provided the NFL with ML the expertise and the workforce to label the unstructured data at scale

Result

Reduced the timeline for developing a high-quality labeling dataset by more than 80%

Using data to construct sophisticated ML models to improve the football experience for its players, teams, and fans

NFL is able to track every player on every play in every NFL game



“Amazon SageMaker Ground Truth was truly a force multiplier in accelerating our project timelines. We wouldn’t have been able to make the strides as quickly as we have without AWS.”

Matthew Davis
VP of IT development



LG AI Research

DEVELOPED FM USING AMAZON SAGEMAKER AI

Challenge

Needed the most efficient ML platform to train their FM, which handles vast amounts of data and large-scale training and inference. Moreover, the project required a high-performance compute infrastructure and the flexibility to train on dozens of terabytes of data

Solution

Built EXAONE—a FM that can be used to transform business processes—using Amazon SageMaker AI, broadening access to AI in various industries such as fashion, manufacturing, research, education, and finance

Result

Developed the EXAONE AI engine in under a year

Supports linear scaling

35% reduction in cost of building AI engine

60% increase in data preparation speed



By using Amazon SageMaker AI's high-performance distributed training infrastructure, researchers can focus solely on model training instead of managing infrastructure."

Kim Seung Hwan

Head of LG AI
Research Vision Lab



AI21 Labs

ACCELERATED FM DEVELOPMENT WITH AMAZON SAGEMAKER AI

Challenge

AI21 Labs builds large language models. These are huge neural networks with tens to hundreds of billions of parameters and training them is a major challenge due to complexity and resource needs

Solution

AI21 Labs recently trained the Jurassic-Grande model with 17 billion parameters using Amazon SageMaker AI. Amazon SageMaker AI made the model training process easier and more efficient while working perfectly with DeepSpeed library

Result

Easier and more efficient model training

Ability to scale the distributed training jobs easily to hundreds of NVIDIA A100 GPUs

Lower inference costs



Because Amazon SageMaker AI handles node failures, restarts elegantly, and orchestrates large distributed runs, the team working on pretraining the model can focus on core tasks.”

Dan Padnos

Vice President of Platform,
AI21 Labs

The AI21 Labs logo, with 'AI21' in white and 'labs' in red, set against a background of a glowing blue and purple particle field.

AT&T Cybersecurity

HELPS BUSINESSES IMPROVE THREAT DETECTION AND
RESPONSE USING MACHINE LEARNING WITH AMAZON SAGEMAKER AI

Challenge

AT&T Cybersecurity wanted to improve threat detection capabilities and deliver a better customer experience with enriched, contextual security notifications and fewer false positives for its customer

Solution

AT&T Cybersecurity off-loaded infrastructure management to AWS and benefited from a scalable solution that offered its customers fast and frequent training for personalized ML models and cost-effective model hosting for inference

Result

50%–60% increase in productivity of data scientists

Improved quality of alert tickets

Smarter, enriched alerts for quick action and reduced analyst time

Optimization of compute costs

Telecommunications | United States

AT&T Cybersecurity is a global managed-security-services provider offering a portfolio of network, security operations, and consulting services aimed at helping to guide and support the current and evolving security needs of organizations



Because Amazon SageMaker AI handles node failures, restarts elegantly, and orchestrates large distributed runs, the team working on pretraining the model can focus on core tasks.”

Dan Padnos

Vice President of Platform,
AI21 Labs



Forethought Technologies

OPTIMIZING COSTS AND PERFORMANCE FOR GENERATIVE AI

Challenge

Forethought Technologies (Forethought) wanted to improve its machine learning (ML) costs and availability as it gained new customers

Solution

Forethought migrated the inference and hosting of ML models to Amazon SageMaker AI, which is used to build, train, and deploy ML models for virtually any use case with fully managed infrastructure, tools, and workflows

Result

80% cost reduction using Amazon SageMaker Serverless Inference

Improved resource efficiency and availability

Improved customer response times and hyper personalization

Software & Internet | United States

Forethought Technologies is a startup in the United States providing a generative artificial intelligence suite for customer service that uses machine learning to transform the customer support life cycle. The company powers over 30 million customer interactions a year

“By migrating to Amazon SageMaker AI Multi-Model Endpoints, we reduced our costs by up to 66 percent while providing better latency and better response times for customers.”

Jad Chamoun

Director of
Core Engineering, Forethought Technologies



EvolutionaryScale

CUTTING-EDGE MODELS CAPABLE OF GENERATING A NOVEL PROTEIN

Challenge

EvolutionaryScale, a frontier AI research lab for biology, needed to empower scientists to advance applications from drug discovery and materials science to carbon capture

Solution

EvolutionaryScale launched a milestone AI model, ESM3, capable of generating novel proteins, which is now available in Amazon SageMaker AI. ESM3 was trained with 1 trillion teraflops – more compute than any other known model in biology – on a dataset of 2.78 billion proteins across the Earth's natural diversity. It is the first generative model for biology that simultaneously reasons over the sequence, structure and function of proteins. This enables scientists to understand and create new proteins, making biology programmable. [Learn more](#)

Result

Sped up a process that would take
500 million years of evolution to occur naturally

Rethinking the possible

EvolutionaryScale makes cutting-edge models capable of generating a novel protein, accelerating a process that would take 500 million years of evolution to occur naturally, made available in Amazon SageMaker AI



EvolutionaryScale



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

perplexity

MOVE QUICKLY AND GENERATE VALUE

Challenge

Faced with the challenge of optimizing its models for accuracy and precision, Perplexity needed a robust solution capable of handling its computational requirements for building one of the world's first conversational answer engines

Solution

Perplexity AI, a company that is currently building one of the world's first conversational answer engines, uses the power of generative AI to help users find relevant knowledge. To elevate the user experience, Perplexity leveraged advanced machine learning infrastructure, training libraries, and inference tools from AWS, Perplexity gained the flexibility, performance, and efficiency required to serve a global user base at scale. [Learn more](#)

Result

Accelerates foundation model training
by 40% with Amazon SageMaker HyperPod

Saving time, effort, and costs

Perplexity AI accelerates foundation model training by 40%
with Amazon SageMaker HyperPod



© 2025, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Ads

AI-powered image generation to help advertisers deliver a better ad experience for customers



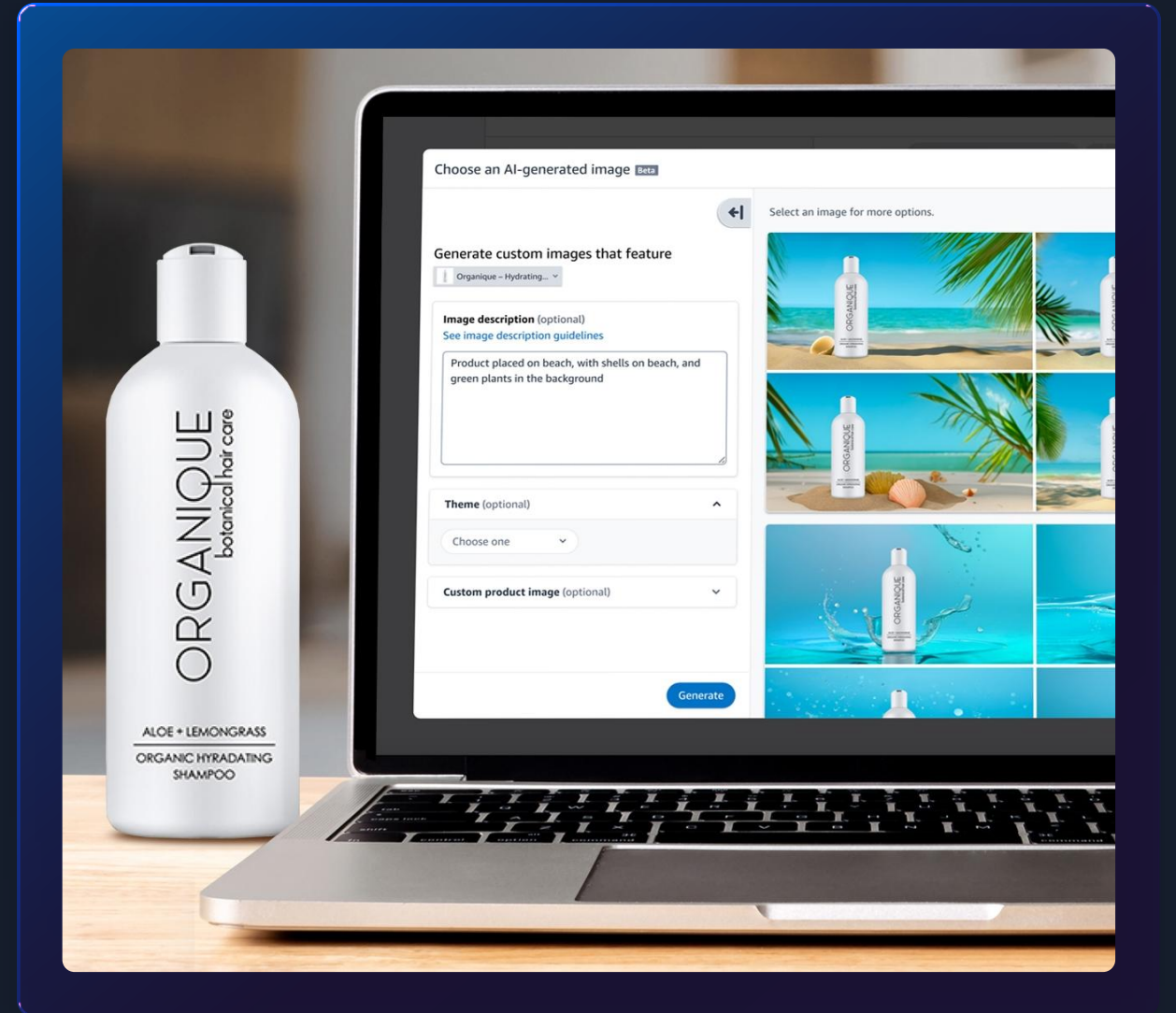
Remove creative barriers and enable brands to produce lifestyle imagery that helps improve their ads' performance



Available to everyone—simply select a product and click generate



Powered by Amazon SageMaker AI



PolyAI

MOVE QUICKLY AND GENERATE VALUE

Challenge

PolyAI needed build customized AI voice assistants for enterprise customers, to customize and scale their contact center support for spoken conversations over the phone

Solution

To offer customized voice AI solutions for enterprises, PolyAI developed natural-sounding text-to-speech models using Amazon SageMaker AI. And they build on Amazon Bedrock to ensure responsible and ethical AI practices. They use Amazon Connect to integrate their voice AI into customer service operations. [Learn more](#)

Result

Able to provide their customers with stable, secure, and scalable voice assistants that are built for enterprises

Customizing Voice AI

PolyAI enables enterprise brands to bring personality to contact centers, using foundation models and Amazon SageMaker AI



Thomson Reuters

MOVE QUICKLY AND GENERATE VALUE

Challenge

Thomson Reuters wanted to explore and innovate across the organization, making AI solutions including generative AI accessible to both technical and nontechnical teams

Solution

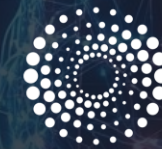
Using AWS, Thomson Reuters built a web-based playground called Open Arena, where users can experiment with a growing set of ML tools powered by large language models (LLMs). Using the platform's chat-based interface, employees without coding backgrounds can explore and develop solutions. One solution is a tax research generative AI application called Checkpoint Edge with CoCounsel. [Learn more](#)

Result

- Accelerating AI model deployment from days to hours
- Streamlining testing and innovation
- Fueling innovation across company

Enabling workforce innovation

Thomson Reuters launched its own LLM playground in under 6 weeks, unlocking company-wide innovation



THOMSON REUTERS



Thank you!

Anton Kartashov

Solutions Architect

Amazon Web Services | WWPS CEE

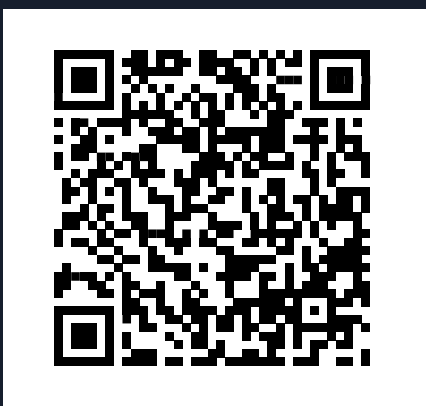
E: antkar@amazon.com

P: +41 76 328 75 44

[linkedin.com/in/anton-kartashov](https://www.linkedin.com/in/anton-kartashov)



Getting Started with SageMaker AI



**Get started with
Amazon SageMaker AI**



**Discover features with
step-by-step tutorials**



**Dive deep with a
hands-on workshop**