



AWS BUILDERS KOREA PROGRAM SPECIAL

SageMaker JumpStart

하루만에 끝내는 생성형 AI

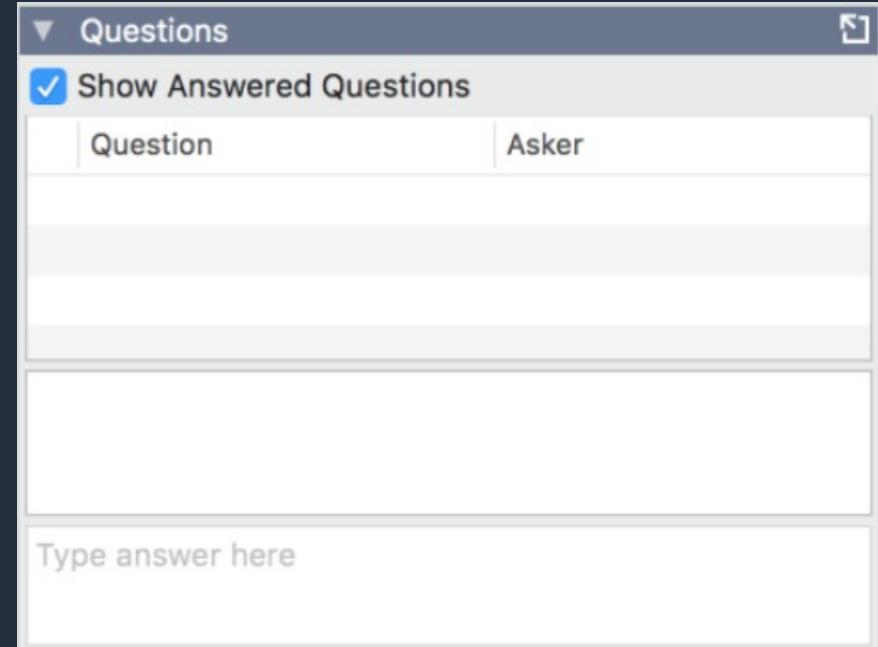
Suji Lee

Solutions Architect
Amazon Web Services

강연 중 질문하는 방법

AWS Builders Korea Go to Webinar “**Questions (질문)**” 창에
자신이 질문한 내역이 표시됩니다. 본인만 답변을 받고 싶으실
경우 (비공개)라고 하고 질문해 주시면 됩니다.

질문 주신 사항에 대해서는 질문창을 통해 답변을 드립니다.



고지 사항 (Disclaimer)

본 컨텐츠는 고객의 편의를 위해 AWS 서비스 설명을 위해 온라인 세미나용으로 별도로 제작, 제공된 것입니다. 만약 AWS 사이트와 컨텐츠 상에서 차이나 불일치가 있을 경우, AWS 사이트(aws.amazon.com)가 우선합니다. 또한 AWS 사이트 상에서 한글 번역문과 영어 원문에 차이나 불일치가 있을 경우(번역의 지체로 인한 경우 등 포함), 영어 원문이 우선합니다.

AWS는 본 컨텐츠에 포함되거나 컨텐츠를 통하여 고객에게 제공된 일체의 정보, 콘텐츠, 자료, 제품(소프트웨어 포함) 또는 서비스를 이용함으로 인하여 발생하는 여하한 종류의 손해에 대하여 어떠한 책임도 지지 아니하며, 이는 직접 손해, 간접 손해, 부수적 손해, 징벌적 손해 및 결과적 손해를 포함하되 이에 한정되지 아니합니다.

실습 시작 전 준비 사항

AWS 계정으로 시작

1. 실습 전 계정을 꼭 신청해주세요 : <https://portal.aws.amazon.com/billing/signup#/start>
2. AWS 계정이 없으신 경우, 행사 참여 전에 미리 AWS 계정 생성 가이드를 확인하시고 AWS 계정을 생성해 주시길 바랍니다.

*AWS 계정 생성 가이드: <https://aws.amazon.com/ko/premiumsupport/knowledge-center/create-and-activate-aws-account/>

3. 검증된 호환성을 위하여 실습 시 사용할 웹 브라우저는 Mozilla Firefox 또는 Google Chrome Browser로 진행 부탁드립니다.

실습 마무리 및 설문 참여 방법

- 실습이 모두 끝난 후에는 **자원 삭제**를 잊지 마세요. 직접 준비하신 AWS 계정으로 실습을 진행하신 고객 분들의 경우, 가이드에 따라 자원 삭제를 진행하셔야 합니다. 또한, 기존에 사용하시던 자원이 있으신 고객 분들의 경우, **오늘 생성한 자원만 삭제하는 것에 주의 부탁드립니다.**
- **가이드:** (세션별 제공)
- 마지막으로 세션이 끝난 후, **GoToWebinar** 창을 종료하면 설문 조사 창이 나옵니다. 이때, **설문 조사를 진행해 주셔야 AWS 크레딧**(1인당 \$50 크레딧, 전체 세션당 1회 제공)을 제공받으실 수 있습니다.

AWS는 고객 피드백을 기반으로 의사 결정을 수행하며 이러한 피드백은 추후에 진행할 세션 방향을 결정합니다.

더 나은 세션을 위하여 여러분들의 소중한 의견을 부탁드립니다.

감사합니다.

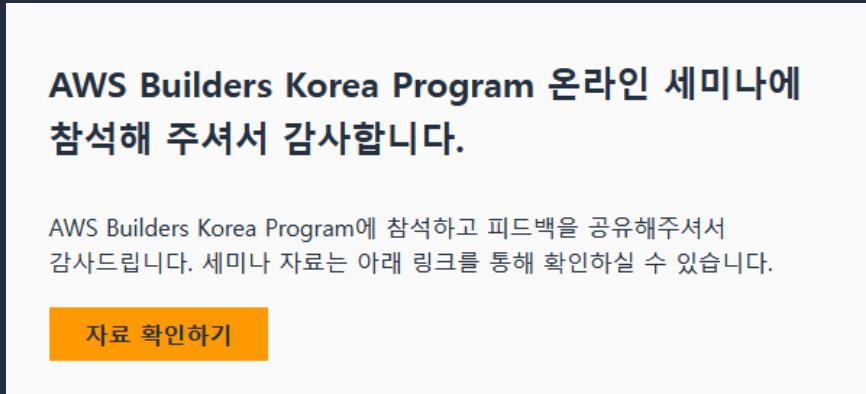
크레딧 안내

- AWS 계정으로 시작하실 경우, 금일 실습에서 발생하는 비용은 당월 과금이 되는 점 미리 확인 부탁 드립니다.
- 웨비나 종료 후 설문 조사에 참여해주신 분들께는 AWS 크레딧 바우처 (1인당 \$50 USD 크레딧, 전체 세션당 1회 제공)를 드립니다.
- 해당 AWS 크레딧은 등록하신 이메일 계정으로 행사 종료 후 1개월 내 발송 드릴 예정이며, 전달 받은 AWS 크레딧은 바로 사용 가능합니다.

감사 메일 & 참석 증명서

- AWS Builders Korea 세션에 참석해 주신 분들께 행사 종료 후 1개월 내 감사메일과 참석 증명서가 순차 발송됩니다.
- 등록 진행 후 참석하지 않으실 경우 별도 메일 및 증명서는 발급되지 않습니다.

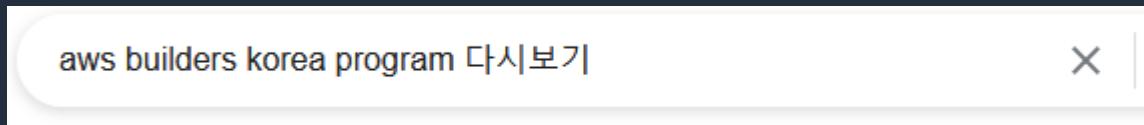
감사 메일 예시



참석 증명서 예시

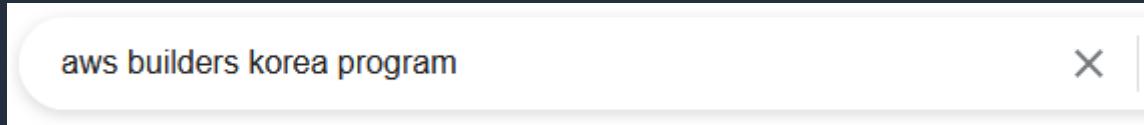


강연 다시보기



<https://kr-resources.awscloud.com/aws-builders-korea-program>

AWS Builders Korea 프로그램 정보



<https://aws.amazon.com/ko/events/seminars/aws-builders/>

SageMaker JumpStart

고객의 과제

ML을 시작하려면 다음 단계가 필요합니다.

too long...

- 공개적으로 사용 가능한 알고리즘 및 모델을 SageMaker로 가져오기
- SageMaker 호환 스크립트 유지 및 업데이트
- 인프라 설정
- NLP 및 비전 모델 처음부터 구축
- 모델 공유 및 협업이 수동으로 수행

Amazon SageMaker JumpStart

몇 번의 클릭만으로
배포할 수 있는 기초
모델, 내장 알고리즘,
사전 구축된 ML
솔루션을 갖춘 ML 허브



머신러닝 허브

사전 학습된 모델, 사전 학습된 기초 모델, 솔루션 및 예제 노트북을 통해 400개 이상의 내장 알고리즘을 찾아보세요.



사전 구축된 훈련 및 추론 스크립트

SageMaker와 호환되며 사용자 정의 데이터 세트로 구성 가능



UI 및 API 기반

단일 클릭 모델 배포를 위한 사용자 인터페이스 또는 Python SDK 기반 워크플로용 API 사용



예제가 포함된 notebooks

전체 ML 워크플로를 안내하는 예시와 함께 선택한 모델을 사용하려면 노트북으로 이동하세요.



조직 내에서 공유 및 공동작업

모델과 노트북을 조직 내 다른 사람들과 공유하고, 그들이 자신의 데이터로 훈련하거나 추론을 위해 있는 그대로 배포할 수 있도록 하세요.

산업 적용 사례



커뮤니케이션

Chatbots, question answering, search



헬스케어

Protein folding, drug development, personalized medicine, improved medical imaging



미디어 & 엔터테이먼트

Video game generation, upscaling content, face synthesis, film preservation & coloring



자동차

Autonomous vehicles, design parts for fuel efficiency



금융 서비스

Risk management, fraud detection



소비자

Optimize pricing and inventory, correctly flag product brand and category



에너지 & 유틸리티

Design renewable energy sources optimized for geo, predictive maintenance



기술 하드웨어

Chip design, robotics

SageMaker JumpStart 에서 foundation models 을 사용하는 이유

1

모델 제공자가 제공하는
기초 모델 선택

AI21labs

Lighton

We bring Light to AI

stability.ai
co:here



alexa

2

모델을 사용해 보거나
배포해 보세요.



Try out models via
AWS Console



Deploy the model for
inference using SageMaker
hosting options includes
single node

3

모델 미세 조정 및
ML 워크플로 자동화



Only selected models
can be fine-tuned



Automate ML
workflow

모델, 인스턴스, 로그,
모델입력, 모델
출력을 포함한
데이터는 계정에
유지됩니다.

Amazon SageMaker
기능과 완전히
통합됨

가장 광범위한 기초 모델 선택으로 구축

AMAZON SAGEMAKER JUMPSTART 에서 이용 가능



Models
Llama 2 7B, 13B, 70B

Tasks
Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation



Models
Jurassic-2 Ultra, Mid
Contextual answers

Tasks
Summarize
Paraphrase
Grammatical error correction

Tasks
Text generation
Long-form generation
Summarization
Paraphrasing
Chat
Information extraction
Question answering
Classification



Models
Cohere
Command XL

Tasks
Text generation
Information extraction
Question answering
Summarization



Models
Falcon-7B, 40B
Open LLaMA

Tasks
RedPajama
MPT-7B, Dolly
BloomZ 176B
Flan T-5 models (8 variants)
DistilGPT2
GPT NeoXT
Bloom models (3 variants)

Tasks
Machine translation
Question answering
Summarization
Annotation
Data generation

Features

Fine-tuning on FLAN T5 models,
GPT-6B, Falcon-7B



Models
Stable Diffusion XL
2.1 base
Upscaling
Inpainting

Tasks
Generate photo-realistic images from text input
Improve quality of generated images

Features

Fine-tuning on Stable Diffusion 2.1 base model



Models
Lyra-Fr
10B, Mini

Tasks
Text generation
Keyword extraction
Information extraction
Question answering
Summarization
Sentiment analysis
Classification



Amazon SageMaker JumpStart를 사용하여 기존 기반 모델 위에 구축



Amazon SageMaker JumpStart is a machine learning (ML) hub that can help you accelerate your ML journey. Explore how you can get started with built-in algorithms with pretrained models from model hubs, pretrained foundation models, and prebuilt solutions to solve common use cases. To get started, see documentation or example notebooks that you can quickly execute.

Reset Filters

Sort By Popularity

Product Type

- Foundation Model
- Model
- Solution

Text Tasks

- End-to-end Solution
- Text Classification
- Text Embedding
- Text Generation
- Text Summarization
- Named Entity Recognition
- Question Answering
- Zero-Shot Classification

Getting started with Amazon SageMaker JumpStart

Products / Machine Learning / Amazon SageMaker JumpStart

Amazon SageMaker JumpStart

Overview Features Pricing FAQs By Role By ML Lifecycle Getting Started Customers Partners

FOUNDATION MODEL FEATURED

Text Generation

Falcon 40B Instruct BF16

Huggingface

Model ID: huggingface-textgeneration-open-llama. This is a Text Generation model built upon a Transformer model from Hugging Face. It is a permissively licensed (Apache-2.0) open source reproduction of Meta AI's LLaMA 7B trained on the RedPajama dataset.

FOUNDATION MODEL FEATURED

Text Generation

Open LLaMa

Huggingface

Model ID: huggingface-textgeneration-open-llama. This is a Text Generation model built upon a Transformer model from Hugging Face. It is a permissively licensed (Apache-2.0) open source reproduction of Meta AI's LLaMA 7B trained on the RedPajama dataset.

FOUNDATION MODEL NEW

Text to Image

Stable Diffusion XL Beta V0.8

StabilityAI

Extend beyond just text-to-image prompting. Stable Diffusion XL offers several ways to modify the images: Inpainting - edit inside the image, Outpainting - extend the image outside of the original boundaries.

FOUNDATION MODEL NEW

Text Generation

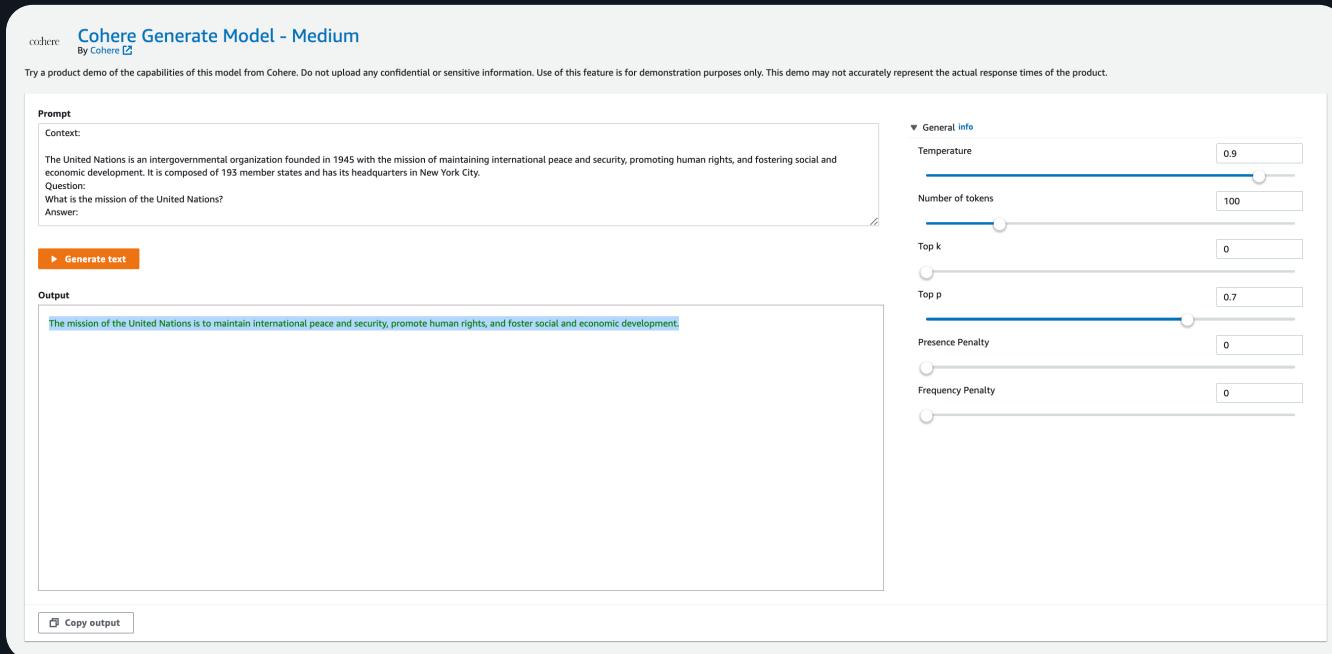
Cohere Command

Cohere

Generative model that responds well with instruction-like prompts. This model provides businesses and enterprises with best quality, performance and accuracy in all generative tasks. And with our intuitive SDK, unlocking the full potential of LLMs for your applications has never been easier.



경험해 보세요



- 코드를 실행하거나 비용을 발생시키지 않고 모델 및 모델 프롬프트를 사용해 보세요.
- HELM 벤치마크 상위 10위의 독점 모델과 비교 목적의 공개 모델에 사용 가능
- 이는 SageMaker 에스크로 계정의 공유 환경입니다.

호스팅에 적합한 인스턴스 선택

Size of model (# of parameters)	Large 3B–10B	Mega 11B–20B	Massive 100B+*
Task Type	Image generation Simple text classification (Short form)	Natural language understanding (NLU)	Natural language generation (NLG) (long form)
Minimum instance required	p3.2xlarge g5.2xlarge	p3.8xlarge g5.12xlarge	p4de.24xlarge p4d.24xlarge
Pricing	\$4/hr \$2/hr	\$15/hr \$9/hr	\$47/hr \$38/hr

Scale vertically (larger instances) to improve latency

Scale horizontally (more instances) to support higher traffic

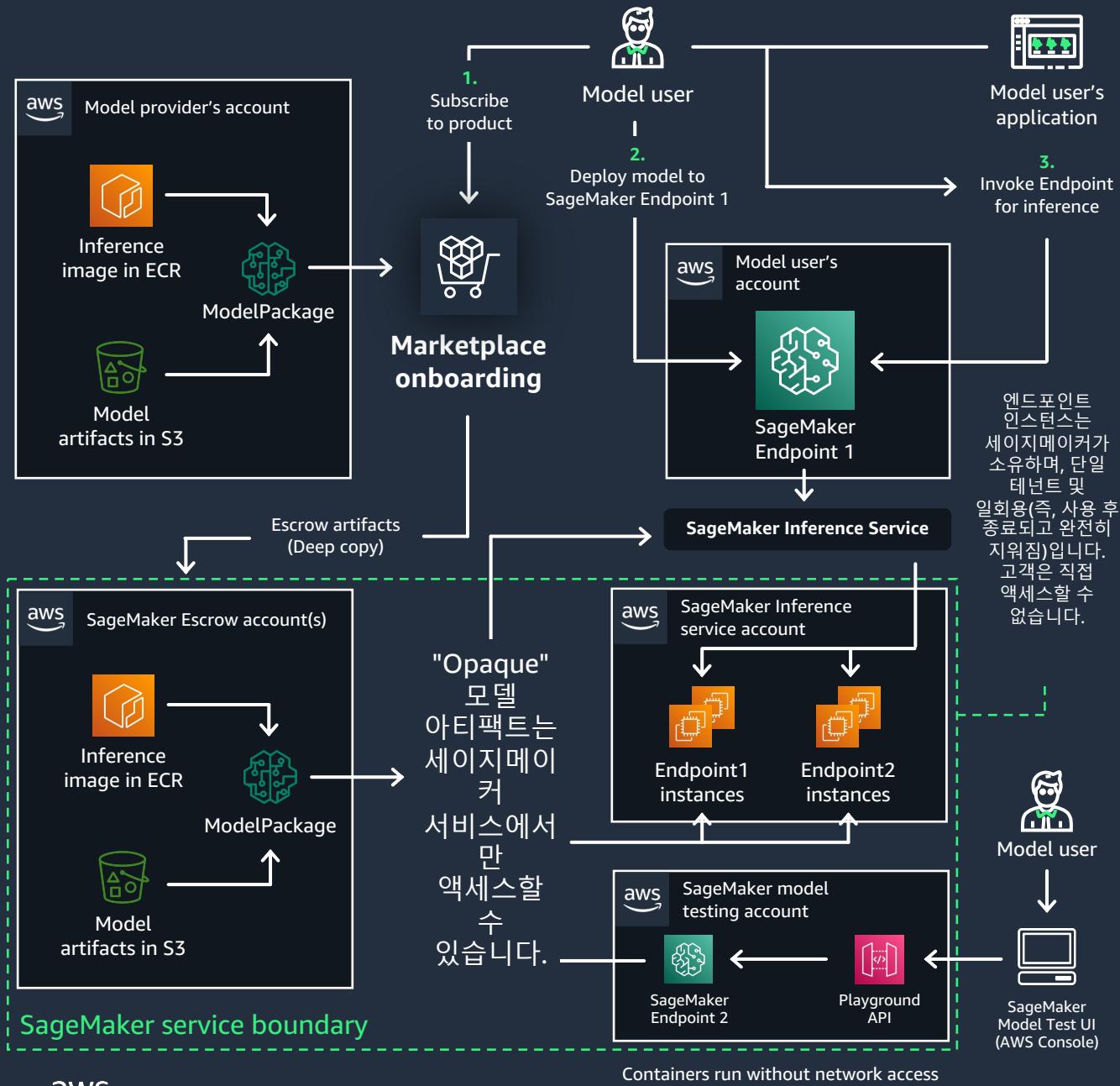
*P4d instances will have limited availability, escalate to S-Team for support

모델에 대한 인스턴스 유형/크기 권장 사항

Model	Instance
Text Generation	
J2 Ultra	g5.48xlarge
J2 Mid	g5.12xlarge
Cohere Command Medium	g5.xlarge
Cohere Command XL	p4d.24xlarge
FLAN T5 XL	g5.2xlarge
FLAN T5 XXL	g5.12xlarge
FLAN UL2	g5.12xlarge
GPT-J 6B	g5.12xlarge
GPT NeoX	g5.24xlarge
Image Generation	
Stable Diffusion 2.1 base	g5.2xlarge
SD Upscaling	g5.2xlarge

SageMaker JumpStart에서 기초 모델을 Fine tune 하기 위한 인스턴스 유형

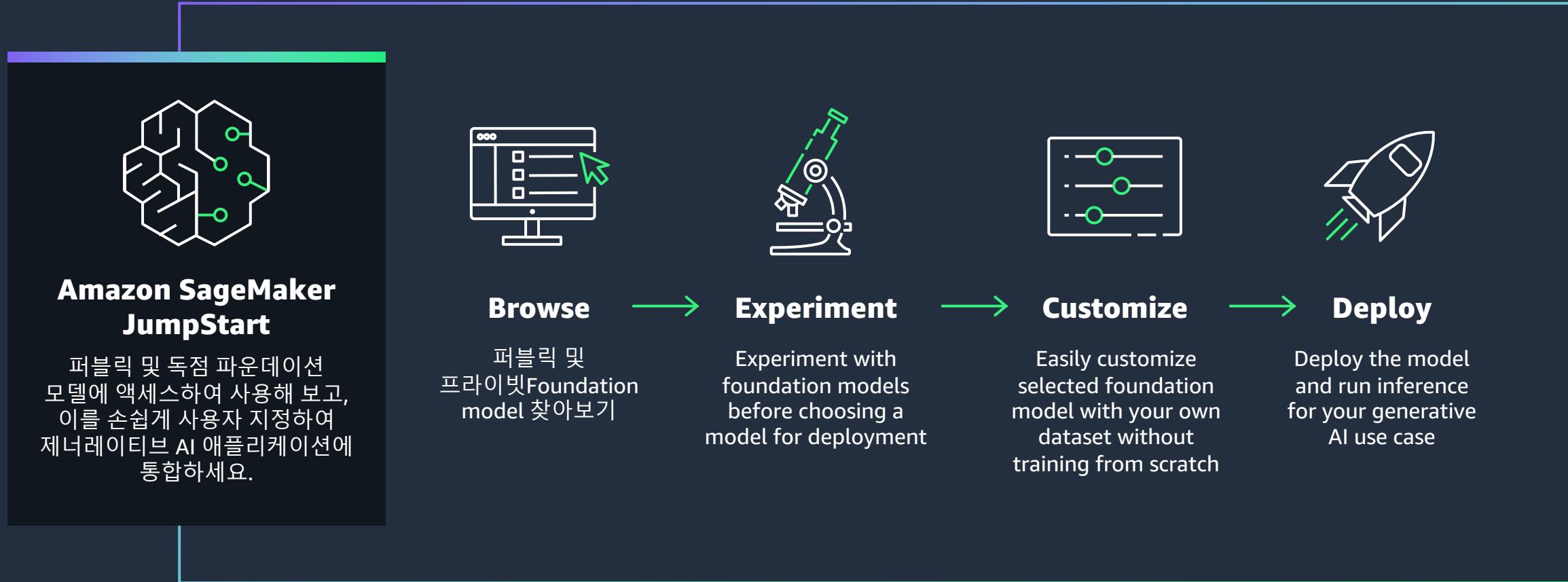
	G4dn	P3	P3dn	P4d	
아이디어	파라미터가 1억 개 미만인 중소규모 모델의 학습 가속화	1억~3억 개의 파라미터로 중대형 모델 훈련 가능 단일 노드 분산 훈련에 적합	300M 이상의 파라미터로 대규모 모델 훈련하기 스팟 트레이닝은 P4d보다 더 나은 가격 대비 성능을 제공할 수 있습니다. 다중 노드 분산 훈련에 적합	클라우드에서 최고의 훈련 성능을 원하는 고객 3억 개 이상의 파라미터가 포함된 대규모 모델 훈련 다중 노드 분산 훈련에 적합	
핵심 기능	16 GB/GPU PCIe only 25–50 Gbps networking 100 Gbps on bare-metal	16 GB/GPU 200–300 GB/s NVLink (4, 8 GPUs) 10–25 Gbps networking	32 GB/GPU 300 GB/s NVLink (8 GPUs) 100 Gbps networking	40 GB/GPU 600 GB/s NVLink (8 GPUs) 400 Gbps networking	
GPU 사양	1, 4, or 8 NVIDIA Tesla T4s	1, 4, or 8 NVIDIA Tesla V100s	8 NVIDIA Tesla V100s	8 NVIDIA Tesla A100s (latest)	
최근 출시	Habana Gaudi	Trainium	g5	p4de*	가장 광범위하고 완벽한 분산 교육 인프라 선택지



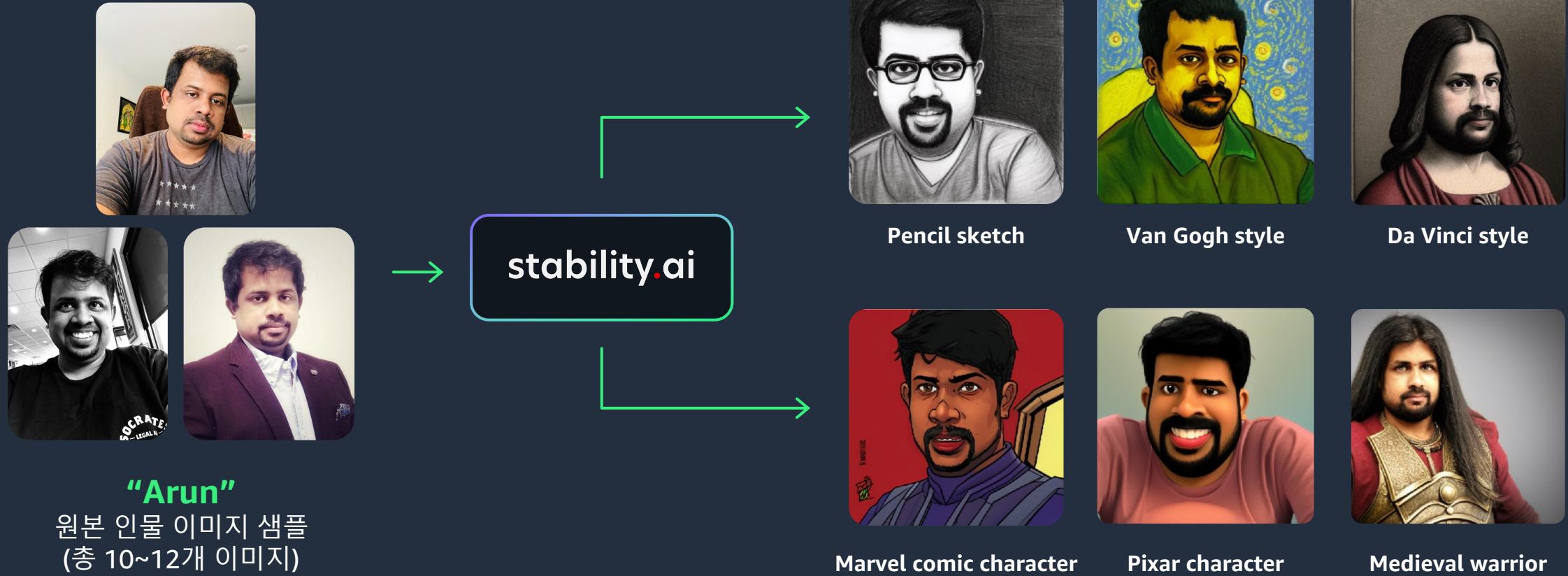
SageMaker JumpStart 는 데이터 및 모델 공급자 IP를 보호합니다.

- 독점 모델 패키지 및 엔드포인트는 세이지메이커 소유 에스크로 계정에서 호스팅됩니다.
- 컨테이너는 아웃바운드 네트워크에 액세스할 수 없으며, 사용자 데이터와 모델 공급자 IP가 동시에 보호됩니다.
- JumpStart가 고객에게 제공하는 기본 모델을 업데이트/트레이닝하는데 데이터가 사용되지 않습니다.

Foundation models: 작동방식



Demo 1: Stable Diffusion fine-tuning 을 사용한 이미지 생성



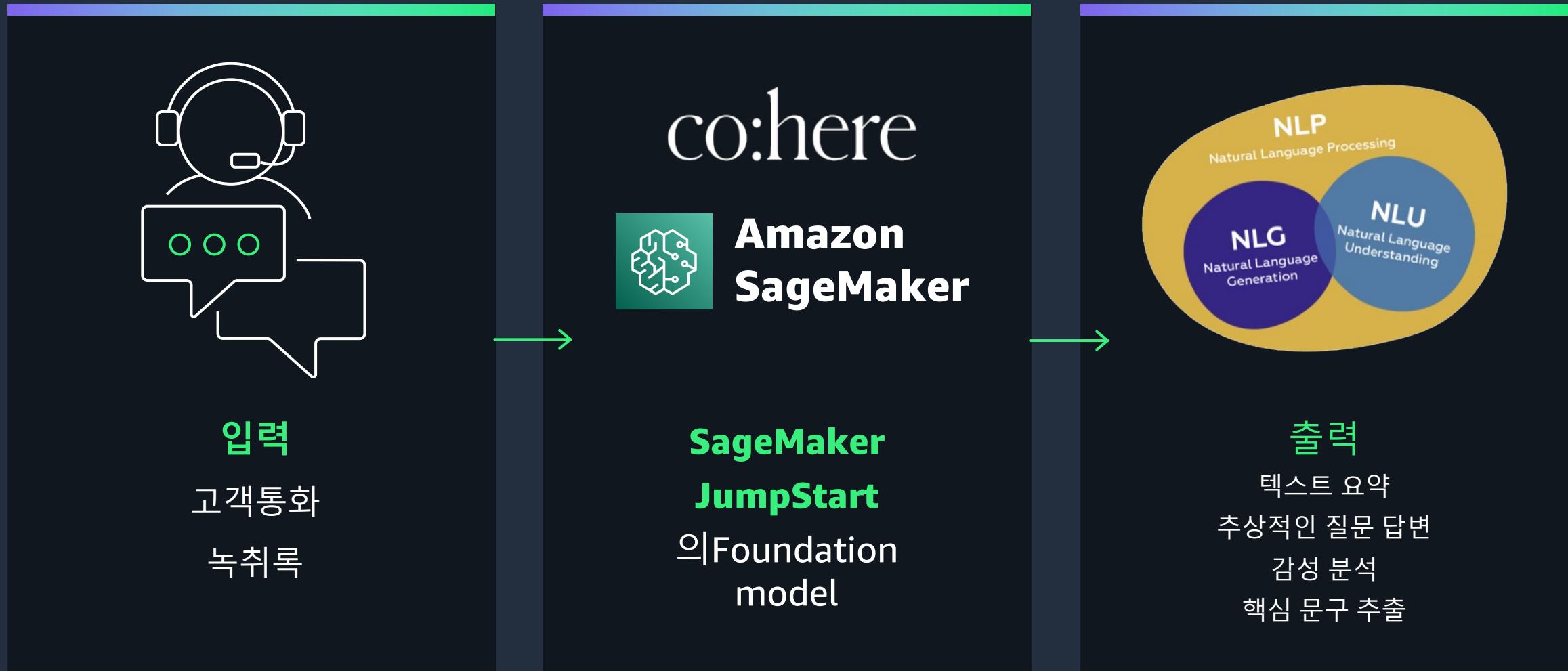
Demo 1: Stable Diffusion fine-tuning 을 사용한 이미지 생성



stability.ai



Demo 2: Co:here Medium 을 사용한 텍스트 생성



예시

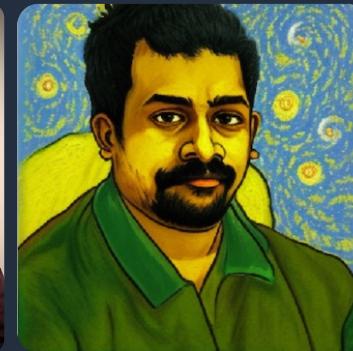
Lab 1

Stable Diffusion
(Getting started)



[LCNC Immersion Day \[link here\]](#)

Fine-tune Stable Diffusion
(나만의 이미지에서)



Lab 2

NLU/NLG with FLAN-T5-XL



[LCNC Immersion Day](#)

NLU/NLG with AlexaTM

AlexaTM 20B

대규모 다국어 Seq2seq 모델을
사용한 Few-shot learning

NLU/NLG with Co:here Medium

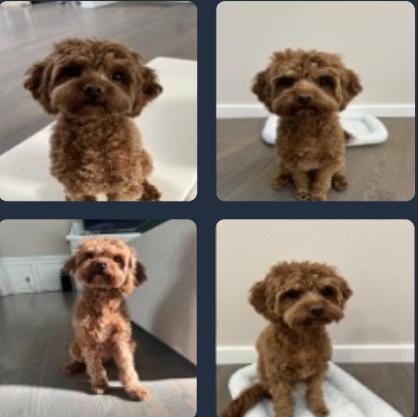
co:here



Amazon
SageMaker

예시

Image generator



학습 데이터셋
도플러 강아지



이미지 생성
도플러의 연필 스케치

Stable Diffusion with fine tuning 으로
이지를 생성합니다.
타겟팅된 콘텐츠 제작(웹 페이지, 사용자
경험 개인화, 리테일 카탈로그 구축)을
시연하는 데 유용하게 사용할 수 있습니다.

Document summarizer

Inputs

Input

The tower is 324 metres (1,063 ft) tall, about the same height as an 81-storey building, and the tallest structure in Paris. Its base is square, measuring 125 metres (410 ft) on each side. It was the first structure to reach a height of 300 metres. Excluding transmitters, the Eiffel Tower is the second tallest free-standing structure in France after the Millau Viaduct.

Summarization Model

Output

Output

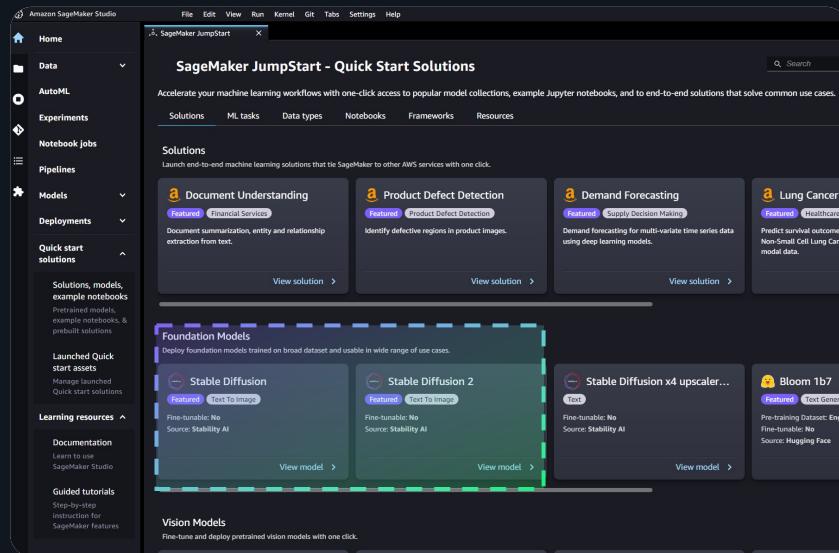
The tower is 324 metres (1,063 ft) tall, about the same height as an 81-storey building. It was the first structure to reach a height of 300 metres.

(대표 이미지)

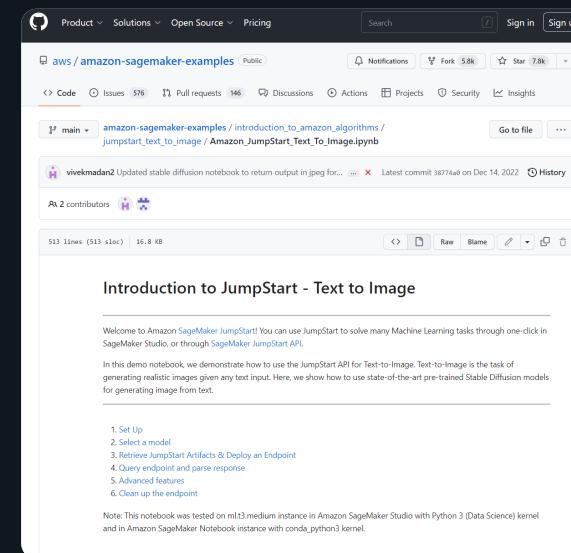
데이터와 모델의 개인정보가
보호되는 환경에서 **텍스트 생성
모델**을 사용하여 여러 문서의
텍스트와 이미지를 동시에
요약합니다.

세이지메이커 Jumpstart로 Foundation Model을 사용하는 3가지 방법

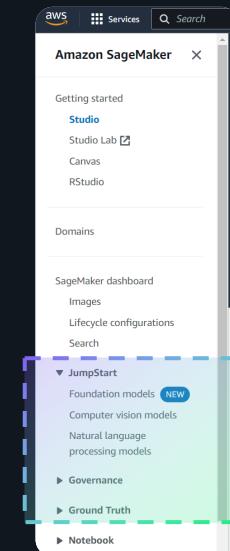
SageMaker Studio One-click deploy



SageMaker Notebooks



AWS console Preview

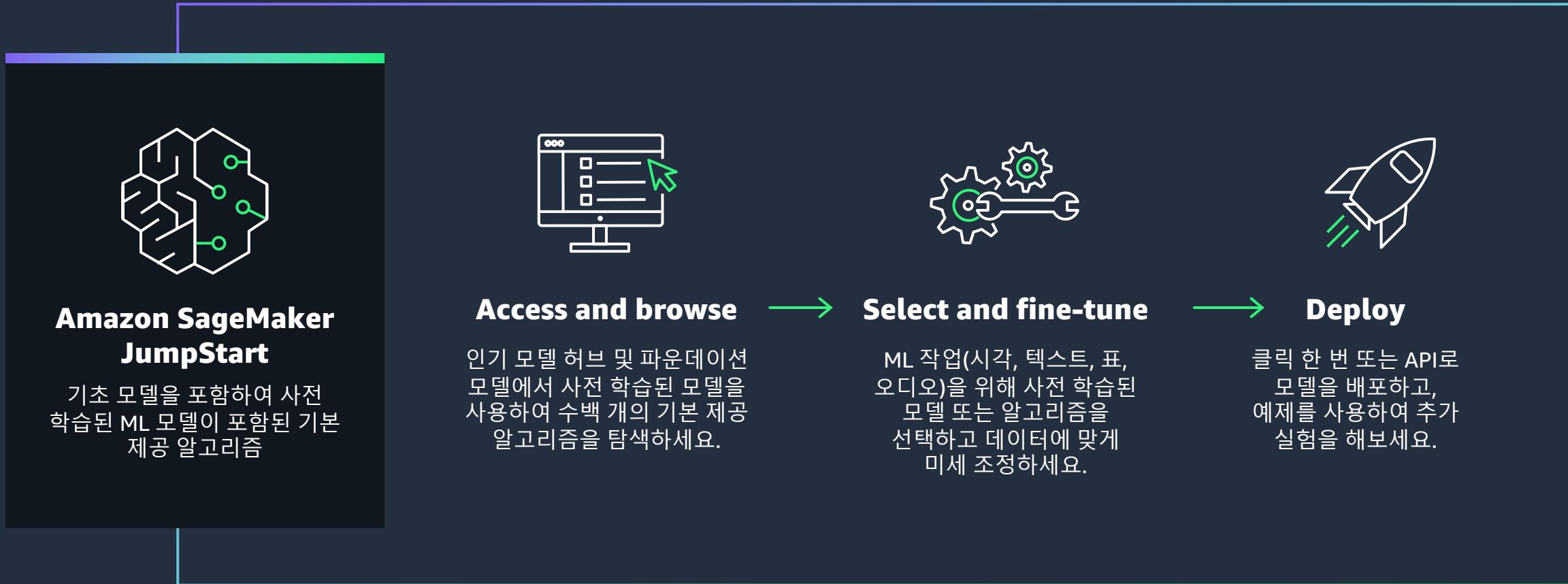


기본 제공 알고리즘 및 사전 학습된 모델

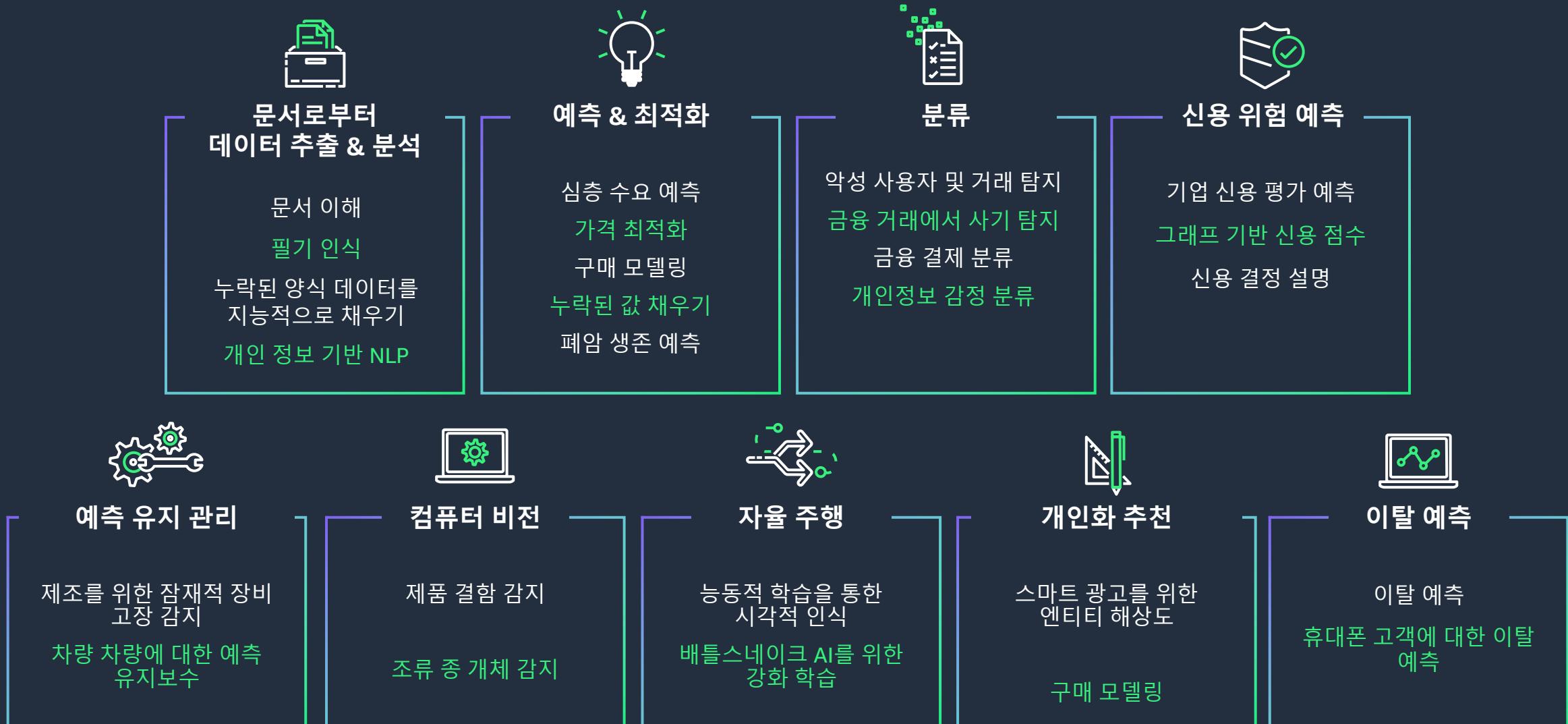
AWS 소유 환경에 안전하게 저장된 PYTORCH HUB, TENSORFLOW HUB, HUGGING FACE HUB의 400개 이상의 알고리즘 및 사전 학습된 최신 공개 모델

	Tasks	Algorithms/models	
 Tabular	Classification, regression, time-series	LightGBM, CatBoost, AutoGluon, TabTransformer, XGBoost, DeepAR	
 Vision	Image classification Image embedding	Object detection Semantic segmentation	ResNet, Inception, MobileNet, SSD, Faster RCNN, YOLO, and more
 Text	Sentence classification Text classification Question answering	Summarization Text generation, translation, Named-entity recognition	AlexaTM, Bloom, Stable Diffusion 2.0, BERT, RoBERTa, DistilBERT, Distillbart xsum, GPT2, ELECTRA, and more
 Audio	Audio embedding	TRILL, TRILLsson, TRILL-Distilled, FRILL	

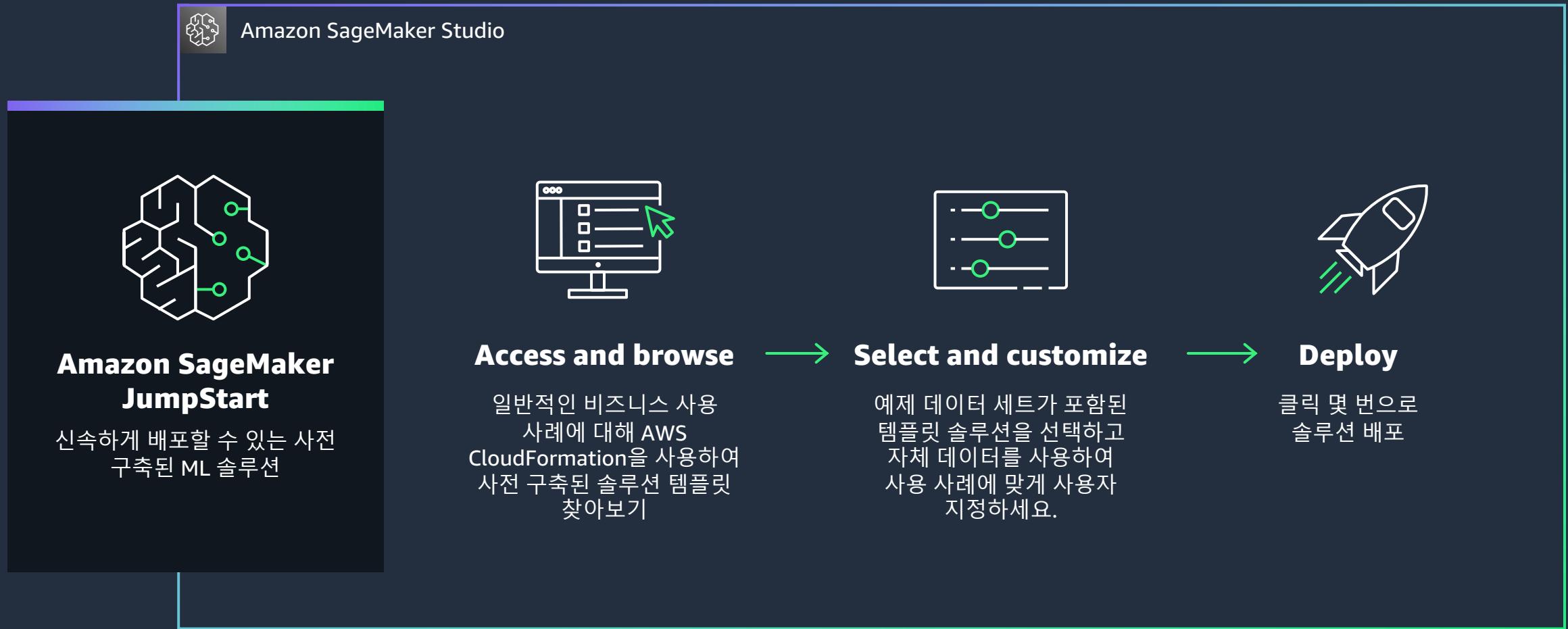
사전 학습된 모델이 포함된 내장 알고리즘: 작동 방식



SageMaker JumpStart 를 사용한 솔루션



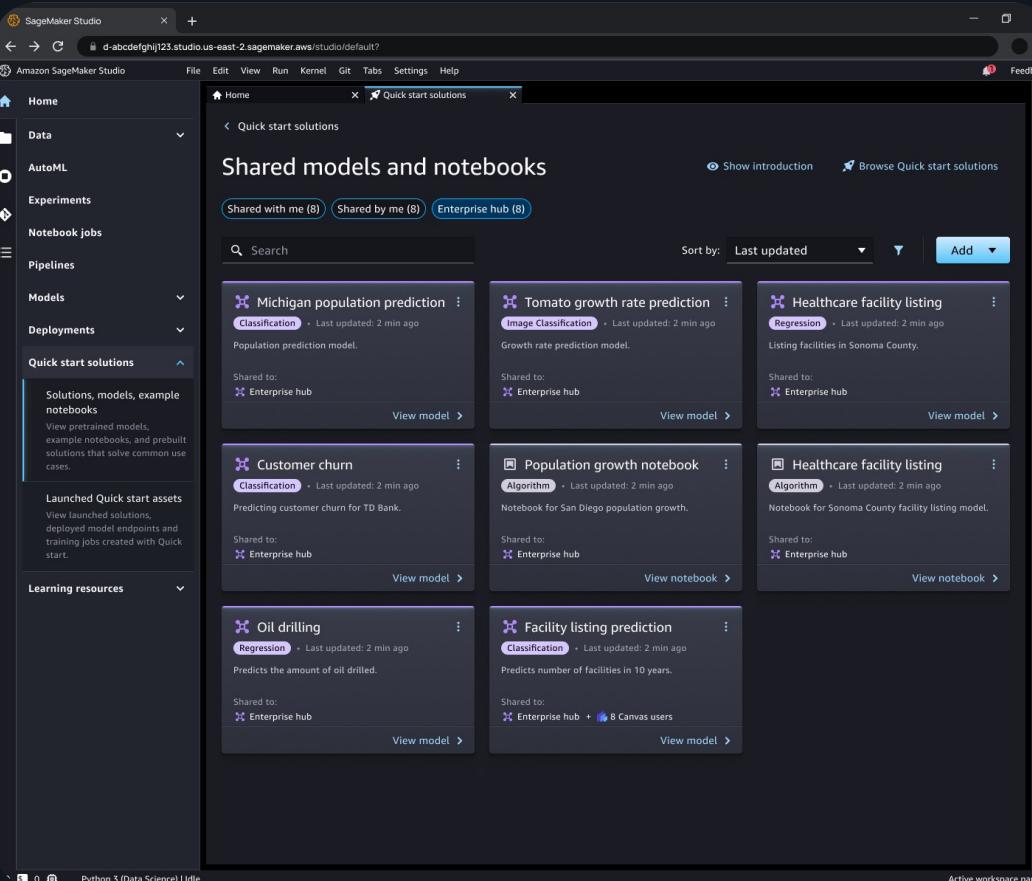
솔루션: 동작 방식



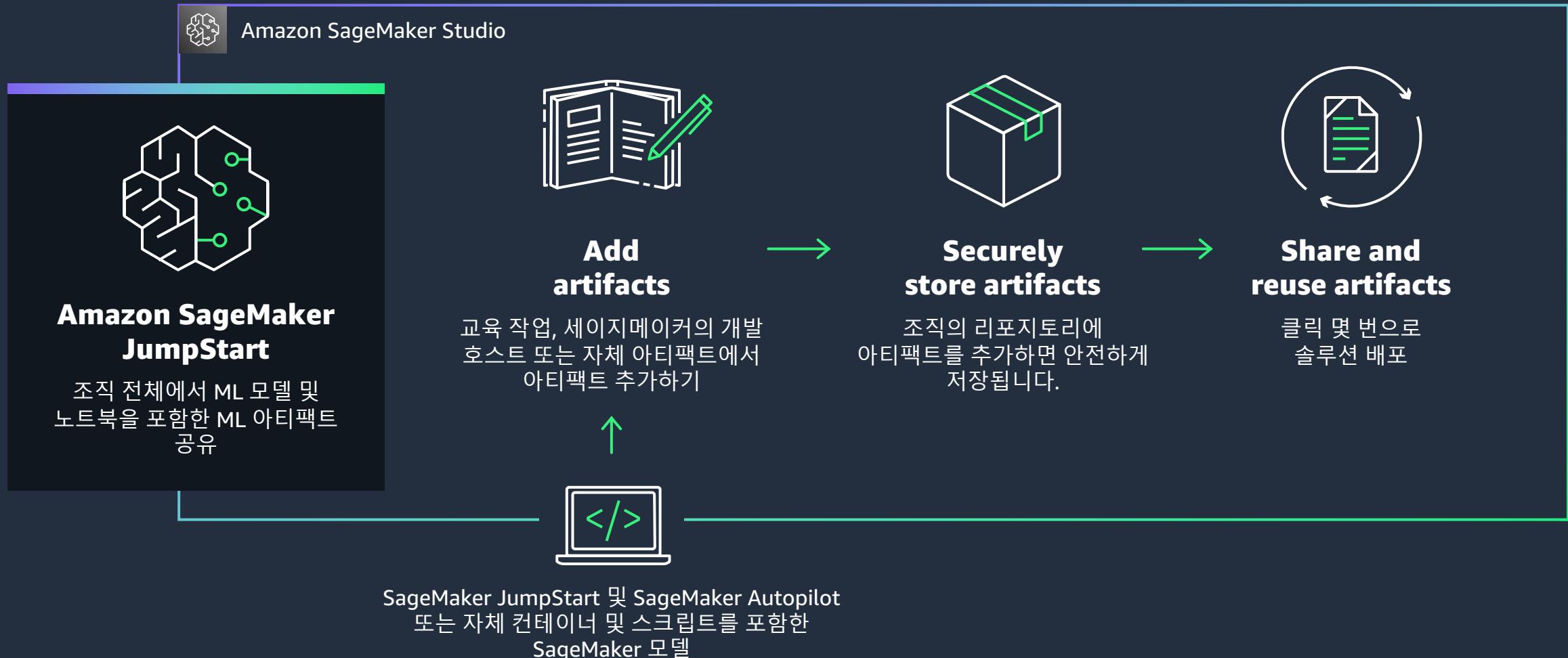
SageMaker JumpStart에서 모델 Artifacts 공유

데이터 과학자가 기업 내에서 ML 아티팩트를 안전하게 공유하고 SageMaker에 내장된 콘텐츠와 함께 재사용할 수 있습니다.

- 조직의 다른 사용자와 공유
- 공유 콘텐츠를 쉽게 검색하고 자체 데이터로 미세 조정을 시작하세요.
- 공유되는 콘텐츠 모니터링 및 제어



ML artifact 공유: 동작 방식



Amazon SageMaker JumpStart 가격 및 가용성



일반적으로
사용 가능



모든 세이지메이커 스튜디오
고객이 **추가 비용 없이 사용 가능**

사용한 리소스에 대해서만 결제 가능



세이지메이커 스튜디오를
사용할 수 있는 모든
지역에서 사용 가능

**(SageMaker Jumpstart
콘솔 환경 미리 보기)**

SageMaker Jumpstart 시작하기



Visit PDP and explore



[JumpStart Product Detail Page](#)
[Wiki: GenAI on SageMaker JumpStart](#)



Engage AWS AI/ML Specialist

Engage your account team
or AI/ML Specialists

1

JumpStart
Hands-on
Workshop

2

Design
Partner/POC
in a Box

3

ML
Solutions
Lab

Demo



Thank you!

Suji Lee

awsjlee@amazon.com