Amazon Bedrock

생성형 AI 애플리케이션을 구축하고 확장하는 가장 쉬운 방법

김기철

Solutions Architect AWS



Agenda

Amazon Bedrock 소개

• Amazon Bedrock 기능

◆ Amazon Nova 소개

Amazon Nova 사용 사례

• 시작하기



Amazon Bedrock 소개

AWS Generative AI Stack

생산성 향상을 위한 애플리케이션



Amazon Q Business

인사이트 및 자동화



Amazon Q Developer

소프트웨어 개발 수명주기

생성형 AI 앱 구축과 확장을 위한 모델과 도구



Amazon Bedrock

아마존 모델 | 파트너사 모델

AI 모델 구축 및 학습을 위한 인프라



Amazon SageMaker

관리형(MANAGED) 인프라



AWS Trainium AWS Inferentia



GPUs

고성능 컴퓨팅

Amazon Bedrock 는 복잡함을 **단순화**합니다



최적의 모델 선택

Amazon 자체 모델(1P) 및 파트너 모델(3P)을 포함한 선도 기업의 고성능 기반 모델(FM)을 활용하세요



비용, 지연 시간, 정확도 최적화

사용 모드, 기능, 그리고 다양한 FM을 선택해 비용, 지연 시간, 정확도 간의 균형을 효과적으로 맞추며 확장하세요



자신의 데이터로 맞춤화

기업의 프라이빗 데이터에서 맥락 정보를 활용해 더 관련성 높고 정확하며 맞춤화된 응답을 제공합니다



안전하고 책임 있는 AI 점검 적용

애플리케이션 요구 사항과 책임 있는 AI 정책에 맞춘 보호 장치를 구현하세요



에이전트 구축 및 오케스트레이션

기업 시스템, API, 데이터 소스와 매끄럽게 연결해 자동화를 실현하세요



Amazon Bedrock

BROAD CHOICE OF MODELS



Coming soon

Anthropic Claude Models on Bedrock

CHOOSE THE EXACT COMBINATION OF INTELLIGENCE, SPEED, AND COST TO SUIT YOUR NEEDS

	*	*		**	***
	Claude 3.7 Sonnet	Claude 3.5 Sonnet	Claude 3 Haiku	Claude 3 Sonnet	Claude 3 Opus
Use case	가장 지능적인 최초의 하이브리드 추론 모델	대규모 처리 작업에 최적화된 설계	가장 낮은 비용으로 가장 빠른 성능 제공	지능, 속도, 비용 간의 완벽한 균형	다양한 형태의 프롬프트를 인간과 같은 이해로 탐색
Context	200k	200k	200K	200K	200K
Vision	✓	√	✓	√	✓
Cost* Input:	\$0.003 \$0.015	\$0.003 \$0.015	\$0.00025 \$0.00125	\$0.003 \$0.015	\$0.015 \$0.075

^{*}Per 1K tokens



Amazon Nova Foundation Models

최첨단 기반 모델들이 최전선의 지능과 업계 선도적인 가격 대비 성능을 제공합니다.

Understanding models

Creative content generation models

Amazon Nova **Micro**

매우 낮은 비용으로 가장 낮은 지연 시간으 응답을 제공하는 텍스트 전용 모델입니다

GENERALLY AVAILABLE

Amazon Nova **Lite**

가벼운 작업에 매우 빠르게 반응하며 가정 저렴한 멀티모달 모델입니다

GENERALLY AVAILABLE

Amazon Nova **Pro**

정확도, 속도, 비용의 최적 조합을 갖춘, 다양한 작업에 적합한 고성능 멀티모달 모델입니다

GENERALLY AVAILABLE

Amazon Nova Premier

복잡한 추론 작업에 가장 뛰어나며, 커스텀 모델을 압축(distill)할 [최고의 교사 모델로 사용되는 가장 강력한 멀티모달 모델입니다

COMING SOON

Amazon Nova Canvas

최첨단 이미지 생성 모델입니다

GENERALLY AVAILABLE

Amazon Nova **Reel**

최첨단 비디오 생성 모델입니다

GENERALLY AVAILABLE

Lower Cost & Latency

Increasing Intelligence

