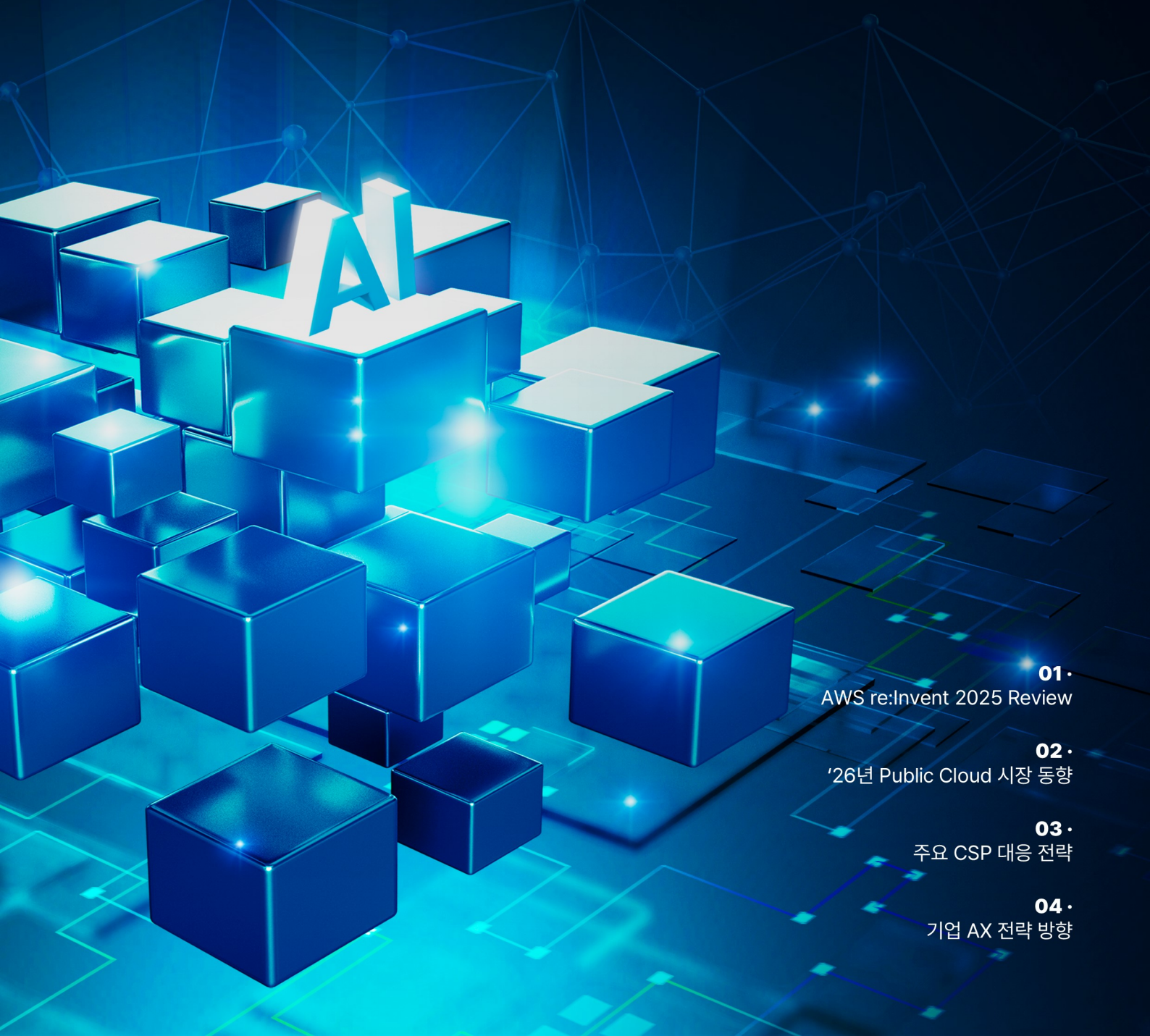
2026·02

SK 주식회사 AX Ackerton Partners

AX 전략 수립

AWS re:Invent Review: Agentic AI 시대 Cloud와 AX 전략

글 | SK AX, CloudEnableX본부 허민회 본부장
SK AX, CloudEnableX본부 이진주 매니저



Executive Summary

Agent at Scale: AI 활용 패러다임 전환

- 지난 12월 진행된 AWS re:Invent는 AI 활용 패러다임 전환에 따라 Public Cloud 산업이 어떤 패러다임으로 진화하고 있는가를 보여주는 행사였다.
- AI/Public Cloud 시장은 '모델 성능 경쟁'을 넘어 '수 억 개의 AI Agent를 사고 없이 운영할 수 있는 구조 경쟁'으로 전환 중이다.

AWS의 지향점? AI/Agent Full Stack Provider

- AWS는 이번 re:Invent에서 스스로를 AI/Agent Full Stack Provider로 재정의 했다.
- 특히, 대규모 Agent를 플랫폼 레벨로 통제하는 'Bedrock Agent Core'를 강조하는 등 AIOps 역량을 드러냈다.

Public Cloud 시장 변화

- AI가 단순 Assistant를 넘어 Agent Workforce로 진화함에 따라 Public Cloud 산업 전반이 재편되고 있다.
- 즉, Cloud 소비 구조 자체가 변화하고 있다. 고객은 Cloud Right 배치 전략으로 이동 중에 있고, Agent 환경에서 비용 폭증과 통제 불가능을 두려워한다.

Public Cloud 시장 변화에 대한 CSP 대응 전략

- AI/Public Cloud 시장이 변화함에 따라 CSP 역시 기업별로 상이한 대응 전략을 이어가고 있다.
- AWS와 MS Azure, Google Cloud Platform의 Agentic AI 전략을 살펴보려고 한다.

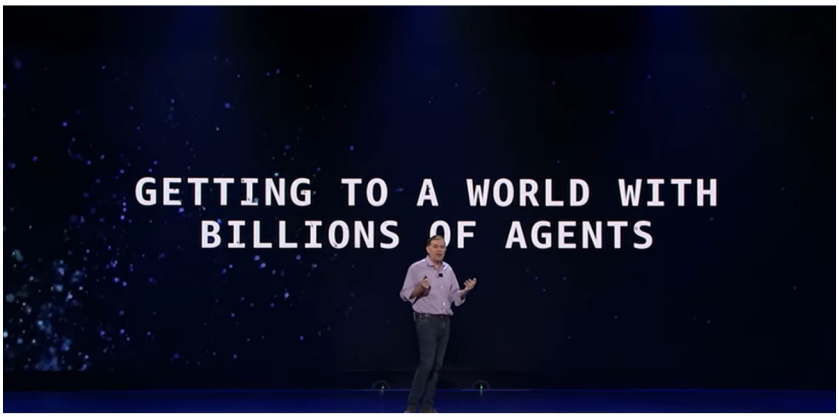
Agent 시대 AX 전략 방향

- 고객이 원하는 것은 기술의 '성능'이 아니라 실제 고객 환경에서 검증된 '지속 가능성'과 '운영 가능성'이다.
- 하나의 뛰어난 Agent 도입이 아닌, Multi Agent가 안정적으로 운영 가능한 AIOps에 대한 고민이 필요하다.

01 AWS re:Invent 2025 Review

지난 12월 현장에서 체감한 **AWS re:Invent**는 단순한 신기능 발표 행사를 넘어 **Public Cloud 산업이 어떤 운영 패러다임으로 진화하고 있는지를 보여주는 자리**였다. AWS의 CEO Matt Garman은 기조연설을 통해 기존의 Cloud at Scale 중심 전략에서 벗어나, 대규모 AI Agent 환경을 전제로 한 Agent at Scale 시대로의 전환을 시사했다. 이는 **Public Cloud의 역할이 '인프라 제공자'에서 'AI 운영 플랫폼'으로 확장되고 있음을 의미**한다.

[그림 1] AWS re:Invent 2025 - Keynote with CEO Matt Garman



출처: YouTube AWS Events

1.1 Agent at Scale: AI 활용 패러다임 전환

이번 re:Invent의 Matt Garman 기조연설을 관통한 메시지는 AI 활용 방식의 근본적인 변화였다.

From Prompt to Action	기존 LLM이 '질문과 답변(Q&A)' 중심이었다면, 이제는 API를 호출하고 외부 시스템과 상호작용하며 실제 결과물을 만들어내는 '액션(Action)' 중심으로 실행 단위가 재편된다.
From Individual to Workforce	Enterprise 운영 환경은 개별 직원의 챗봇 수준을 넘어, 수만 개의 에이전트가 서로 협력하며 비즈니스 프로세스를 완결 짓는 '에이전트 워크포스(Agent Workforce)'로 진보하고 있다.

[표 1] Assistant vs Agent vs Workforce

	AI (현재)	AI Agent (가까운 미래)	Agent Workforce (예상되는 미래)
역할	질문 응답, 요약, 보조	업무 수행, 판단, 실행(Action)	조직 단위 업무 자동화/운영
실행 단위	단일 요청 처리	다수 Action의 연쇄	수백만~수십억 Action 동시 실행
업무 범위	개인 단위, 단일 Task	기능/프로세스 단위	전사/조직 단위
지성	일회성 (Stateless)	상태 기반 (Stateful)	지속 실행
Risk	비교적 낮고 단순	비용/오작동 Risk	비용 폭증/보안/통제실패 Risk

왜 Scale이 핵심인가

AWS가 주목한 것은 Agent 자체가 아니라, **Scale이 가져오는 불확실성**이다. Agent 수가 급증할수록 고객의 비용/보안 Risk와 복잡성이 증가한다. 잘못 설계된 자동화는 운영 효율이 아니라 장애/비용 폭증을 불러온다. 즉, **Agent 시대의 경쟁력은 AI 모델 성능이 아니라, Multi Agent를 통제/운영/중단할 수 있는 구조와 AIOps 역량에 있다.**

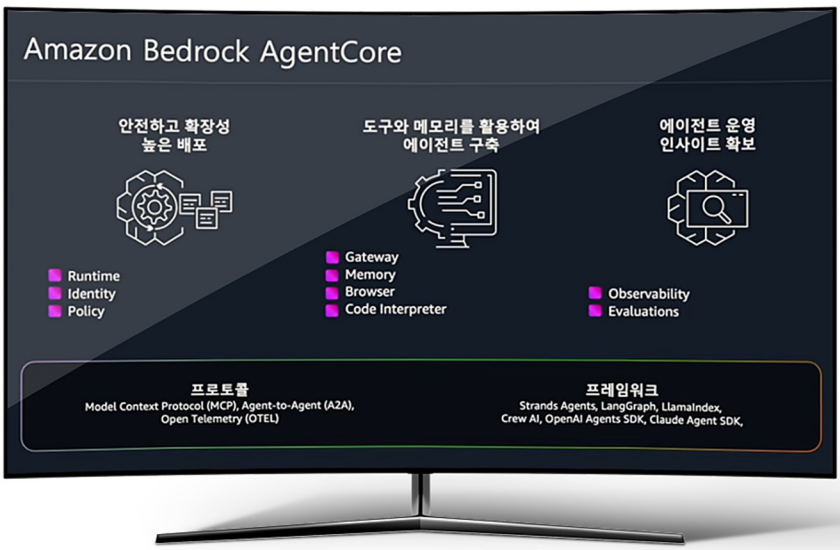
1.2 AWS의 전략적 포지셔닝

AI/Agent Full-stack Provider

AWS는 대규모 Agent 환경에서 AI/Agent의 전체 Life Cycle을 지원하는 Full Stack Provider로 자리매김하겠다는 의지를 드러냈다. 인프라/플랫폼/보안/거버넌스/요금제 등 모든 도구를 AI/Agent 위주로 개편했고, 고객의 End-to-End AI/Agent 여정을 지원한다는 방침이다.

기존 re:Invent의 혁신이 무엇을 더 빠르고 저렴하게 제공하는 것에 초점을 맞췄다면, 올해 행사는 AWS가 Multi Agent를 실제 운영 가능한 구조로 만드는 것에 얼마나 진심인지가 돋보였다. 가장 특징적인 서비스는 'Bedrock AgentCore'로 대규모 Agent 운영 레이어를 서비스화하여 출시했다.

[그림 2] Amazon Bedrock AgentCore 소개



출처: AWS

AWS re:Invent 2025 신기능 발표 요약

올 해 신기능 발표는 단순 'AI 기능 추가'를 나아가, 대규모 Agent 실행을 전제로 Cloud 서비스 전반을 재설계하는 데 집중되었다.

[표 2] re:Invent 2025 AI 관련 신기능 발표

	변화 방향	관련 서비스
AI Agent 라인업 강화	코드/보안/운영/업무 프로세스 모든 과정의 에이전트 서비스 출시	Kiro, Security Agent, DevOps Agent, Bedrock AgentCore, Amazon Quick Suite 등
AI 비용 최적화	단순 GPU 단가 개선이 아닌, AI 서비스 전체 운영 단가 구조 재설계 (요금제 개편, 저장 단가 절감, 인덱싱 비용 절감, 경량 모델 출시 등)	DB Savings Plan, OpenSearch GPU Indexing, Nova2 Lite 등
AI 인프라 혁신	Agent/AI 대규모 실행 가능한 인프라 혁신 (AI 학습/추론 최적화, AI 프라이빗 리전, 장기 서버리스 옵션 등)	Trainium 3 Ultra/4, AI Factories, Lambda Durable Functions 등
AI 파트너쉽 강화	Multi Modal/Platform 생태계 개방 강화, 기업/ISV/스타트업 연합 강화	Bedrock 모델 추가 (Gemma, Mistral 등), 산업별 표준 Agent Blueprint 출시 등

출처: AWS

더 많은 내용을 보시려면

파일 다운받기

버튼을 눌러주세요