디지털서비스

# 이슈리포트

Digitalservice Issue Report

# 워크플로우를 가진 SaaS만 살아남는다

NIA 한국지능정보사회진흥원 디지털서비스 이용지원시스템

2026 Vol. 07

디지털서비스 이슈리포트

# CONTENTS

Digitalservice Issue Report

## 디지털서비스 이슈리포트


01 **워크플로우를 가진 SaaS만 살아남는다** 3

| 김영욱 SAP Product Engineering Product Expert

02 **영국의 CTP 지정과 EU의 DMA 게이트키퍼 예비 지정** 13

| 한상기 테크프론티어

03 **토큰 이코노미시대의 AI 판옵스 전략** 19

| 윤대균 아주대학교

04 **2026년 클라우드 솔루션 리포트: IAM/ID 플랫폼** 27

| 정채상 KAIST

# 01 워크플로우를 가진 SaaS만 살아남는다

| 김영욱 SAP Product Engineering Product Expert

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.


## 이제 대부분의 SaaS는 데이터 저장소일 뿐

2026년 상반기, 이스라엘의 암호화폐 관리 소프트웨어 스타트업 Utila는 회사의 소프트웨어 예산을 50% 삭감했다. 방법은 단순했다. B2B GTM 및 세일즈, 마케팅 영역에서 매우 빠르게 성장 중인 Clay, Vendelux를 비롯한 10개의 SaaS 앱을 해지하고, 잠재 고객 추적부터 이메일 발송, 마케팅 캠페인 관리까지 이 앱들이 수행하던 모든 기능을 AI 에이전트 하나로 통합했다. 그리고 허브스팟만 남겼다. 단, 그 용도는 '데이터 저장소'로 쓰일 뿐이다. 이 변화는 2026년 기업 소프트웨어 시장에서 진행 중인 구조 전환의 본질을 가장 선명하게 압축한다.[1)]

프랑스 제약 기업 사노피(Sanofi)는 앤트로픽 클로드 코드와 스타트업 소프트웨어로 구축한 자체 AI 에이전트를 통해 서비스나우 사용을 줄이고, 아웃소싱 업체가 처리하던 구매 주문 감사 업무를 대체했다. 스타벅스는 마이크로소프트와 IBM 제품의 AI 대안을 모색하고 있다. 업종도, 규모도, 지역도 다른 이 기업들이 동일한 방향으로 움직이고 있다는 사실은, 이것이 일시적 유행이 아니라 구조적 전환임을 가리킨다.

시장은 이미 이 전환의 의미를 인식했다. 2026년 2월, S&P 500 소프트웨어·서비스 지수에서 불과 이틀 만에 2조 달러가 증발했다. 이 사건을 시장은 '사스포칼립스(SaaSpocalypse)'라 명명했다. 촉발 요인은 앤트로픽의 엔터프라이즈 에이전트 플러그인 출시와 세일즈포스, 서비스나우, 구글의 에이전트 제품 출시가 겹친 것이었지만, 그 이면에는 하나의 구조적 인식이 있었다. **AI 에이전트가 워크플로우를 자율적으로 수행하기 시작하면, 사람이 대시보드에 로그인해야 가치가 발생하던 SaaS의 가격 공식이 근본적으로 흔들린다는 것이다.** 가트너는 이를 "사용자 성장과 매출 성장 사이의 연결 고리가 끊어졌다"고 표현했다.[2)]

1) The Information, "Traditional SaaS Loses in Corporate Budget Shift", Jul 9, 2026
2) ciodive.com, "Agentic AI to disrupt $234B in SaaS spending: Gartner", Jul 6, 2026

숫자는 이미 방향을 가리키고 있다. 2026년 1분기 AI 네이티브 기업 지출은 전년 대비 94% 급증한 반면, 전통 SaaS 성장률은 8%에 머물렀다. 가트너는 에이전트 AI가 2030년까지 2,340억 달러 규모의 기업 소프트웨어 지출을 잠식할 것으로 전망하며, 딜로이트는 2026년 기업 디지털 혁신 예산의 50% 이상이 AI 자동화에 배분될 것으로 예측했다.[3] 이 수치들이 공통으로 가리키는 것은 하나다. 20년간 기업 IT 예산의 핵심 항목이었던 SaaS 구독 모델이 구조적 압박에 직면해 있다는 것이다.

이번 글에서는 그 압박의 구조를 분석한다. 핵심 논지는 명확하다. **AI 에이전트 시대에 기업 소프트웨어의 생존을 결정하는 기준은 기능의 풍부함이나 사용자 경험의 완성도가 아니다. 그것은 '워크플로우를 소유하고 있는가, 아니면 데이터를 보관하고 있는가'의 차이다.** 에이전트가 API를 통해 워크플로우를 직접 실행하는 세계에서, 실행 레이어를 잃은 소프트웨어는 데이터 창고로 전락한다. 허브스팟의 운명이 그것을 예고하고 있다.

## 1. 사스포칼립스 임계점은 이미 지났다

'사스포칼립스'를 단순한 주가 조정으로 읽는 것은 본질을 놓치는 것이다. 이것은 새로운 위협이 등장한 날이 아니라, 이미 진행 중이던 구조 전환이 임계점을 넘었다는 시장의 인식이 가시화된 날이다. 투자자들이 실제로 정가에 반영하기 시작한 질문은 하나였다. **AI 에이전트가 사용자 대신 워크플로우를 실행하는 세계에서, 사용자 로그인을 전제로 설계된 소프트웨어의 현재 가격은 정당한가?** 세일즈포스와 서비스나우는 모두 직전 분기까지 시장 기대치를 상회하는 실적을 보고한 상태였다는 점에서, 이 하락은 실적 부진이 아니라 비즈니스 모델의 구조적 취약성에 대한 선제적 재평가였다. 이 재평가의 논리는 명확하다.

AI 에이전트는 RPA라고 불리는 기존 오토메이션과 근본적으로 다른 방식으로 소프트웨어와 상호작용한다. RPA가 화면의 버튼을 클릭하고 정해진 절차를 따르는 것이라면, AI 에이전트는 API를 통해 시스템에 직접 접근해 비정형 입력을 처리하고 다단계 의사결정을 자율적으로 수행한다. 에이전트가 사람 대신 소프트웨어에 접속하는 순간, 그 소프트웨어에 정교한 사용자 인터페이스가 존재하는지 여부는 더 이상 가치 판단의 기준이 되지 않는다. 수십억 달러의 R&D가 투자된 대시보드 디자인과 UX 최적화는 에이전트 앞에서 무의미해진다.

시장의 반응은 예고였고, 실제 기업 예산 데이터는 그 예고가 현실임을 확인한다. AI 네이티브 기업의 94% 대비 전통 SaaS 성장률 8%, 이 두 숫자의 격차는 단순한 신기술 채택 속도 차이가 아니다. 기업

3) 딜로이트, "SaaS meets AI agents", Nov 18, 2025

IT 예산이라는 한정된 파이 안에서 예산 항목의 구조적 이동이 발생하고 있다는 신호다. 이 이동의 특징은 AI 예산이 기존 IT 예산과는 다른 항목에서 나오고 있다는 점이다. 딜로이트 리포트는, 기업들이 AI 자동화에 투입하는 예산의 상당 부분은 소프트웨어 항목이 아니라 운영 지출(OpEx) 항목, 더 구체적으로는 인건비 항목에서 전환되고 있다고 분석한다. 기업들은 소프트웨어와 인력이라는 두 축을 동시에 AI로 대체하기 시작했다.

한편 이 전환에서 수혜를 받는 소프트웨어도 존재한다. 사노피의 사례에서 서비스나우 사용은 줄었지만 SAP 사용은 오히려 늘었다. AI 에이전트 투자가 확대됨에도 불구하고 SAP와 같은 ERP 시스템의 위상이 오히려 강화되는 현상은, 이 전환이 소프트웨어 전체의 소멸이 아니라 소프트웨어 내부에서의 선별적 재편임을 시사한다.

## 2. "에이전트 접근세" – 벤더들의 방어 전략과 그 한계

'사용자 수×좌석 단가'가 매출이 되는 전통적인 Per-Seat 모델의 붕괴를 인식한 소프트웨어 벤더들이 꺼내든 방어 전략이 있다. 플랫폼 밖의 AI 에이전트가 자사 데이터에 접근할 때 별도의 과금을 부과하는 것이다. 이를 '에이전트 접근세(Agent Access Tax)'라고 한다. 허브스팟이 투자자들에게 예고한 바로 그 전략이다. 외부 AI 에이전트가 자사 CRM 데이터를 호출할 경우 과금하겠다는 계획은, 사용자 로그인이 줄어도 데이터 접근 자체를 수익화하겠다는 논리다. 서비스나우도 유사한 방향을 검토하고 있으며, SAP와 오라클은 이미 '간접 접근' 과금 정책을 통해 이 전략을 만들어 놓은 상태다.[4)]

단기적으로 이 전략은 매출 방어 수단으로서 일정한 효과를 가질 수 있다. 데이터 소유권이 있는 플랫폼은 에이전트가 그 데이터를 필요로 하는 한 협상력을 유지한다. 그러나 이 전략이 내포하는 구조적 역설은 분명하다. 에이전트 접근에 과도한 비용을 부과하는 벤더는 고객에게 플랫폼 이탈의 경제적 동기를 제공한다. Utila의 책임자는 허브스팟이 외부 에이전트에 과금하더라도 계속 사용하겠다고 밝혔지만, 그 이유는 그 워크플로우 가치 때문이 아니라 이미 축적된 데이터를 이전하는 비용이 더 크기 때문이다. 데이터 이전 비용이라는 락인이 사라지는 순간, 에이전트 접근세는 고객 이탈을 가속화하는 촉매가 될 것이다. 기업들은 이미 이 위험을 인식하고 소프트웨어 계약 갱신 시점에 에이전트 접근 조건을 명시적으로 협상 의제로 올리기 시작했다. "에이전트가 우리 데이터를 읽고 쓸 때의 비용 구조는 무엇인가? 외부 에이전트와 벤더 자체 에이전트 사이에 차등 요금이 있는가? API 호출량 급증 시

4) Forbes, "AI Agents Could Break SaaS Pricing", May 29, 2026

서차지가 발생하는가?" 이 질문들은 2026년 엔터프라이즈 소프트웨어 구매 협상의 표준 의제가 되고 있다. 계약 이전에 이 조건을 확정하지 못한 기업들은, 에이전트 배포가 본격화되는 시점에 예상치 못한 비용 구조와 마주칠 위험이 있다.

에이전트 접근세 전략이 궁극적으로 직면하는 한계는 더 근본적인 지점에 있다. **데이터를 볼모로 한 방어 전략은 데이터 소유권이 지속되는 동안만 유효하다는 점이다.** 기업들이 데이터를 독립 레이크하우스로 이전하거나, AI 에이전트가 복수의 소스에서 데이터를 재구성하는 역량을 갖추게 되면, 단일 플랫폼의 데이터 독점력은 약화된다. 이 시나리오에서 에이전트 접근세는 단기 수익 방어 수단으로 기능했다가, 플랫폼 전환 비용이 낮아지는 시점에 고객 이탈 트리거로 전환될 수 있다. 방어 전략이 동시에 탈출구를 알리는 신호가 되는 역설이다.

## 3. 기업 예산 이동의 실체

글로벌 제약 기업 사노피의 AI 에이전트 구축은 소프트웨어 구독, 파트너 소프트웨어, 그리고 인력 아웃소싱이라는 세 가지 비용 항목을 동시에 재편했다. 여기에서 눈여겨봐야 할 점은 단순한 비용 절감이 아니라 그 선택의 구조다. 서비스나우는 줄였지만 SAP 사용은 늘렸다. 이 비대칭적 결과는 AI 에이전트 시대의 소프트웨어 생존 법칙을 압축한다. 서비스나우가 수행하던 IT 서비스 관리 워크플로우는 에이전트가 직접 실행 가능한 영역이었다. 반면 SAP가 보유한 ERP 데이터, 즉 구매 주문, 재무 기록, 공급망 정보는 에이전트가 작동하기 위해 반드시 접속해야 하는 '진실의 단일 원천'이다. 에이전트는 SAP를 대체한 것이 아니라 SAP 위에서 작동한다. 이것이 같은 AI 에이전트 도입이라는 의사결정이 두 소프트웨어에 정반대의 결과를 가져온 이유다.

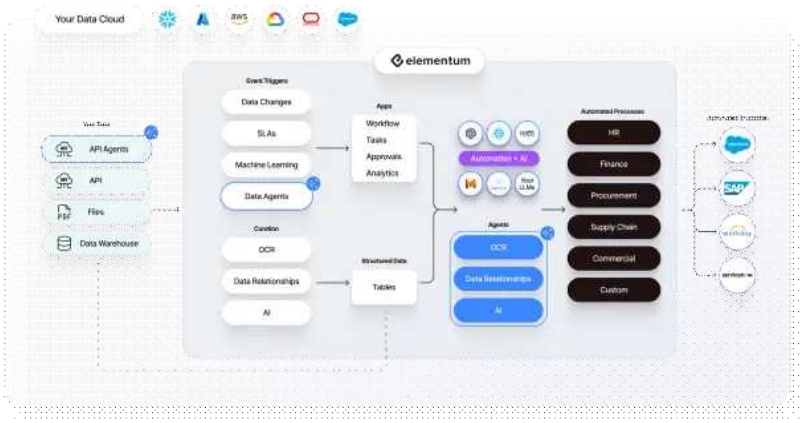


그림 1 사노피가 서비스나우를 대체하기 위해 사용한 것으로 알려진 엘레멘텀(elementum.ai)

이 변화는 기업들이 AI 도입을 계기로 오랫동안 쌓인 레거시 기술 부채를 직시하고, 소프트웨어 스택을 단순화하는 방향으로 전략을 재편하고 있음을 보여준다. 또한 AI를 단순히 기존 소프트웨어에 추가하는 방식이 아니라, 기존 소프트웨어를 AI로 대체하는 방식으로 접근하고 있다. 그리고 그 대체의 대상은 '더 많은 기능을 가진 소프트웨어'가 아니라 '워크플로우의 실행 역할을 담당하던 소프트웨어'다. 반면 데이터를 소유한 소프트웨어는 남는다. 기능이 아니라 데이터 소유권이 생존을 결정하는 새로운 질서가 기업 구매자들의 실제 의사결정 안에서 이미 작동하고 있는 것이다.

사노피의 예에서 또한 주목해야 할 항목이 있다. 바로 인도 아웃소싱 업체의 역할 축소가 그것이다. 이 변화는 소프트웨어 시장의 재편을 넘어, 지난 40년간 글로벌 기업 운영의 핵심 원리였던 노동 차익거래(Labor Arbitrage) 모델의 구조적 종언을 예고한다. 노동 차익거래의 논리는 단순했다. 선진국 기업이 반복적이고 규모가 크며 노동 집약적인 업무를 임금 수준이 낮은 국가의 인력에게 위탁함으로써 비용을 절감하는 것이다. 인도는 이 모델의 최대 수혜자였다. 수백만 명의 중산층이 이 모델 위에서 형성됐다. 이 모델을 작동시킨 조건들이 지금 AI 에이전트에 의해 정면으로 해체되고 있다. 반복성, 예측 가능성, 규모의 경제. 아웃소싱이 유효했던 이유였던 이 세 가지 특성은, 동시에 AI 에이전트 자동화의 최우선 대상이 되는 조건이기도 하다. 오라클은 2026년 상반기 인도에서 1만 2,000명을 감축하며 AI 투자를 확대했고, TCS도 1만 2,000명 감원을 단행했다. 인도 주요 IT 기업들은 2026 회계연도 첫 9개월간 순 고용 증가가 사실상 17명에 그쳤다. 수만 명 단위로 신규 채용하던 산업이 사실상 채용을 멈춘 것이다.[5)]

이 전환이 한국 IT 서비스 산업에 갖는 함의는 직접적이다. 글로벌 기업들이 아웃소싱 대신 AI 에이전트를 선택하는 구조적 흐름은, 대형 SI 프로젝트의 인력 기반 수익 모델에도 동일한 질문을 던진다. 사람이 수행하던 반복 업무를 AI 에이전트가 대체하는 속도가 빨라질수록, 인력 규모를 기반으로 한 프로젝트 과금 방식은 구조적 압박을 받는다. 이 압박은 위협이기도 하지만, 동시에 새로운 역할로의 전환 기회이기도 하다. 에이전트를 구축하고 기업 시스템과 연결하며 안전하게 오케스트레이션하는 역량, 즉 사노피가 외부에 의뢰한 것을 내부화할 수 있도록 돕는 역할이 그것이다. 이 기회를 누가 먼저 구조화하느냐가 향후의 경쟁 우위를 결정할 것이다.

## 4. 워크플로우를 가진 SaaS만 살아남는다.

### 시스템 오브 레코드 vs. 시스템 오브 인게이지먼트

5) Outsource Accelerator, “AI threatens millions of BPO jobs in India and Philippines”, May 6, 2026

기업용 소프트웨어는 오래전부터 두 가지 역할로 구분되어 왔다. 시스템 오브 레코드(System of Record)와 시스템 오브 인게이지먼트(System of Engagement)다. 전자는 데이터의 단일 원천이다. ERP, CRM, HRIS처럼 기업의 핵심 데이터가 생성되고 저장되며 참조되는 플랫폼이다. 후자는 그 데이터를 기반으로 사용자가 실제 업무를 수행하는 인터페이스다. 이메일, 협업 도구, 마케팅 자동화 플랫폼, 프로젝트 관리 앱들이 여기 속한다. 두 레이어는 오랫동안 공존했다. 데이터는 레코드 시스템에 있고, 사람은 인게이지먼트 시스템을 통해 그 데이터와 상호작용했다.

AI 에이전트의 등장은 이 공존 구조를 근본적으로 흔든다. 에이전트는 인게이지먼트 레이어를 필요로 하지 않는다. 목표가 주어지면 레코드 시스템의 API에 직접 접속해 데이터를 읽고, 판단하고, 결과를 기록한다. 사람이 인게이지먼트 도구를 통해 수행하던 중간 단계들이 에이전트의 자율 실행으로 압축된다. 이 구조에서 인게이지먼트 레이어는 에이전트가 우회할 수 있는 선택적 인터페이스로 격하된다. 반면 레코드 시스템은 에이전트가 반드시 접속해야 하는 데이터의 원천으로서 오히려 위상이 강화된다. 이것이 생존의 분기점이다.

소프트웨어가 '워크플로우의 실행자'로 기능해왔는지, 아니면 '데이터의 소유자'로 기능해왔는지가 AI 에이전트 시대의 운명을 결정한다. 전자는 에이전트에게 역할을 빼앗기고, 후자는 에이전트의 필수 접속 대상이 된다. 그리고 이 분기점은 소프트웨어의 브랜드나 시장 점유율, 혹은 지난 10년간 쌓아온 고객 관계와 무관하게 작동한다. 시장에서 가장 오랫동안 사랑받아 온 소프트웨어라도, 그것이 워크플로우 실행 역할에 의존해왔다면 에이전트 시대의 구조적 압박에서 자유롭지 않다. 프로세스 편의성 위에 구축된 도구는 에이전트가 동일한 결과를 API 호출로 직접 달성하는 순간 존재 이유를 잃는다. 이 분기점을 이해하면, 사노피가 서비스나우를 줄이고 SAP를 늘린 선택의 논리가 명확해진다.

## SAP가 살아남는 이유, 허브스팟이 "데이터 창고"로 남는 이유

SAP의 생존 논리는 역설적으로 단순하다. SAP는 기업의 구매 주문, 재무 데이터, 공급망 기록, 인사 정보를 보유한 단일 원천이다. AI 에이전트가 기업 운영과 관련된 어떤 업무를 수행하던, 그 업무의 대부분은 SAP 데이터를 참조하거나 SAP에 결과를 기록하는 과정을 포함한다. 에이전트는 SAP를 대체하지 않는다. SAP 위에서 작동한다. 에이전트 시대에 SAP의 가치는 줄어드는 것이 아니라, 에이전트 오케스트레이션의 허브로서 오히려 강화된다. 수십 년에 걸쳐 기업의 핵심 트랜잭션 데이터를 축적해 온 이 플랫폼들은, AI 에이전트가 확산될수록 더 많은 에이전트 접속 요청을 받게 된다. '에이전트 접근세' 논쟁이 가장 현실적으로 적용되는 것도 이 레이어다.

반면 마케팅 자동화, 이메일 캠페인, 영업 파이프라인 관리, 고객 지원 티켓 처리라는 다층적 워크플로우를 하나의 플랫폼 안에서 실행하는 '올인원 CRM·마케팅 플랫폼'으로 성장해 온 허브스팟은 다른 경로를 걷고 있다. 이 포지셔닝의 핵심 가치는 워크플로우 실행의 통합성이었다. 그러나 AI 에이전트는 이메일 발송, 잠재 고객 추적, 캠페인 관리를 인터페이스를 거치지 않고 직접 수행할 수 있다. 남은 것은 데이터다. 고객 정보, 거래 이력, 커뮤니케이션 기록. 에이전트가 참조해야 할 데이터가 안에 있는 한, 허브스팟은 해지되지 않는다. 그러나 그 역할은 '워크플로우 플랫폼'에서 '데이터 창고'로 축소된다.

## 5. 깃허브(GitHub)의 위기

### 삼면이 포위된 형국

2026년 7월, 깃허브는 전례 없는 경쟁 압박에 직면해 있다. 코드 저장소와 AI 코딩 보조라는 두 개의 핵심 사업 영역이 동시에 공격받고 있다. 커서는 VS Code를 포크해 AI 네이티브 코드 에디터라는 새로운 카테고리를 창조하며 깃허브 코파일럿의 AI 코딩 시장 선점 지위를 잠식해 왔다. 오픈AI는 전 깃허브 임원을 영입하여 코드 저장소 시장에 진출하겠다는 전략적 선언을 했다. 주간 활성 사용자 9억 명의 챗GPT와 개발자 생태계를 보유한 오픈AI가 코드 저장소 시장에 진입하면, 깃허브는 AI 코딩 어시스턴트와 저장소라는 두 사업 모두에서 오픈AI와 정면으로 경쟁하게 된다.

여기에 덧붙여 가장 상징적인 경쟁자가 등장했다. 전 깃허브 CEO가 창업한 스타트업 Entire가 독자적인 깃(Git) 서비스를 출시한 것이다. 깃허브를 가장 잘 아는 사람이 그 대체재를 제공하고 있다. 엔타이어 전략은 깃허브의 가장 취약한 지점을 정확히 겨냥한다. 깃허브는 6월 한 달에만 여섯 차례의 서비스 장애를 기록했는데 AI 에이전트 트래픽의 급증이 주요 원인이다. 사람이 코드를 작성하고 커밋하던 시대와 달리, AI 에이전트가 24시간 자율적으로 코드를 생성하고 수정하며 PR을 올리는 환경에서 트래픽 패턴 자체가 바뀌었다. 깃허브가 자체 서버에서 마이크로소프트 애저로 인프라를 이전하며 장애를 줄이겠다고 밝혔지만, 그 전환이 완료되기 전까지 경쟁자들은 이 공백을 파고들고 있다.

### 마이크로소프트 수직통합 해자에 생긴 균열

애저 클라우드, 깃헙, M365, 팀즈, 오픈AI 파트너십이 하나의 수직통합 스택을 형성하며, 엔터프라이즈 IT 조달 채널 자체가 마이크로소프트 솔루션 안에 있다는 구조는 독립 AI 스타트업에게는 넘기 어려운

진입 장벽이었다. 그리고 지금까지 그 전략의 핵심 사례가 깃허브 코파일럿이었다. 포춘 100 기업 90%의 침투율, 유료 구독자 470만 명, M365 번들 안에 포함되는 조달 구조. 이것이 지금까지의 현실이었다.

하지만 2026년 6월 이후 이 해자에 균열이 생겼다. 수직통합의 강점은 여전히 유효하다. 그러나 수직통합이 전제하는 조건, 즉, 각 레이어가 에이전트 시대의 트래픽 패턴에 적응할 수 있어야 한다는 조건이 충족되지 못할 때 균열이 발생한다는 것이 깃허브의 사례가 드러내는 새로운 법칙이다. **수직통합은 구조적 강점이지만, 그 강점이 유지되려면 각 레이어가 새로운 환경의 요구에 지속적으로 대응해야 한다.** 깃허브의 반복적 장애는 수직통합 구조 안에서도 인프라 레이어가 AI 에이전트 시대의 사용 패턴 변화를 따라가지 못할 때 어떤 일이 벌어지는지를 보여주는 구체적 사례다.

더 구조적인 문제는 오픈AI와의 관계에서 비롯된다. 마이크로소프트는 오픈AI에 수백억 달러를 투자하고 애저를 통해 오픈AI의 모델을 배포하는 핵심 파트너다. 이 파트너십은 AI 전략에서 모델 역량을 확보하는 핵심 수단이었다. 그러나 오픈AI가 깃허브의 저장소 사업을 직접 겨냥하는 경쟁자로 나서는 순간, 마이크로소프트는 자신의 가장 중요한 AI 파트너가 자신의 핵심 자산을 위협하는 역설적 상황에 놓인다. 이 균열이 수직통합 전체의 붕괴를 의미하지는 않는다. 포춘 100 기업들의 마이크로소프트 의존도는 단기간에 바뀌지 않는다. 그러나 깃허브의 사례는 중요한 교훈을 남긴다. 수직통합이 만드는 해자는 정적인 구조물이 아니다. AI 에이전트가 만들어내는 새로운 트래픽 패턴, 새로운 사용 방식, 새로운 경쟁 구도 앞에서, 해자의 각 레이어는 지속적으로 적응해야 한다. 적응하지 못하는 레이어에는 균열이 생기고, 그 균열은 경쟁자들의 진입 공간이 된다.

## 6. 리더십을 위한 함의

### 워크플로우 지도를 다시 구축하라

이 변화에 맞는 전략적 의사결정의 출발점은 하나다. 지금 우리 조직이 사용하는 소프트웨어 포트폴리오를 '워크플로우 실행자'와 '데이터 소유자'로 구분하는 작업이다. 이것은 IT 부서의 시스템 감사가 아니라, 조직 리더가 직접 개입해야 하는 전략적 분류다. 왜냐하면 이 분류의 결과가 향후의 소프트웨어 계약 구조, 아웃소싱 전략, 그리고 AI 에이전트 투자 우선순위를 동시에 결정하기 때문이다.

분류의 기준은 단순하다. 각 소프트웨어에 대해 두 가지 질문을 던지면 된다.

**첫째, 이 소프트웨어의 핵심 가치는 사람이 그 안에서 작업을 수행할 때 발생하는가, 아니면 그 안에 저장된 데이터가 다른 시스템이나 에이전트에 의해 참조될 때 발생하는가?**

**둘째, AI 에이전트가 이 소프트웨어의 API를 직접 호출해 동일한 결과를 만들어낼 수 있는가?**

첫 번째 질문에 "데이터"라고 답하고 두 번째 질문에 "아니오"라고 답한다면, 그 소프트웨어는 에이전트 시대에도 생존한다. 반대의 경우라면, 그것은 교체 또는 역할 축소의 후보다. Utila가 10개 SaaS를 해지하면서 허브스팟만 남긴 의사결정의 논리가 정확히 이 기준을 따랐다.

이 분류가 완료된 후의 행동 우선순위는 세 가지로 정리된다.

첫째, 워크플로우 실행자로 분류된 소프트웨어의 갱신 계약에서는 협상력을 최대한 행사해야 한다. 이미 대안이 존재하고, 벤더들도 이를 알고 있다. 시장 협상력의 균형이 구매자 쪽으로 이동한 시점이다. 둘째, 데이터 소유자로 분류된 소프트웨어에 대해서는 에이전트 접근 조건을 지금 당장 계약에 명시해야 한다. 외부 에이전트의 API 접근 비용, 볼륨 기반 추가 요금, 벤더 자체 에이전트와의 차등 요금 문제를 계약 갱신 이전에 확정하지 않으면, 에이전트 배포가 본격화되는 시점에 예상치 못한 비용 구조와 마주하게 된다. 셋째, AI 에이전트 투자는 워크플로우 실행자를 대체하는 방향에서 시작해야 한다. 사노피가 서비스나우 사용량을 줄이는 동시에 구매 주문 감사 업무를 에이전트로 전환한 순서가 좋은 사례다.

한 가지 함정을 경고해야 한다. 이 전환의 논리를 '비용 절감'으로만 프레이밍하는 것이다. Utila가 소프트웨어 예산 50%를 절감했다는 수치는 인상적이지만, 그것은 결과이지 목표가 아니다. 진짜 전략적 질문은 이것이다. 에이전트가 워크플로우를 실행하는 세계에서, 우리 조직의 핵심 데이터는 어디에 있어야 하며, 그 데이터 위에서 에이전트가 생성하는 결과물의 품질과 신뢰성을 어떻게 보장할 것인가? 비용은 이 질문에 제대로 답한 결과로 따라오는 것이다.

## 한국 SI의 역할 재정의: 에이전트 오케스트레이터로의 변화

한국의 대형 SI 기업들이 지난 30년간 구축해 온 핵심 역량은 이기종 시스템의 통합이었다. SAP, 오라클, 세일즈포스, 자체 개발 레거시 시스템을 하나의 운영 환경으로 연결하고, 그 위에서 기업의 업무 프로세스가 작동하도록 설계하는 것이 SI의 존재 이유였다. 이 역량은 수십 년의 현장 경험, 대형 엔터프라이즈 고객과의 관계, 그리고 인력 기반의 프로젝트 실행력 위에 구축됐다.

AI 에이전트 시대가 요구하는 역량은 이 기반 위에서 확장되지만, 그 방향은 근본적으로 다르다. 시스템을 연결하는 것이 아니라, 에이전트가 시스템 데이터를 안전하게 읽고 쓰며 자율적으로 판단할 수 있도록 오케스트레이션하는 것이다. 어떤 데이터를 어떤 권한으로 에이전트에게 허용할지, 에이전트의 판단 결과를 어떻게 검증하고 기록할지, 복수의 에이전트가 협업할 때 충돌과 오류를 어떻게 관리할지에 대한 아키텍처 설계 역량이 그것이다. 이것이 에이전트 오케스트레이터의 역할이다.

이 역할 전환이 한국 SI에게 위협인 동시에 기회인 이유는 명확하다. 위협의 측면에서, 인력 규모를 기반으로 한 프로젝트 과금 모델은 구조적 압박을 받는다. AI 에이전트가 반복적 업무를 대체하는 속도가 빨라질수록, 사람 수에 비례한 용역 단가는 협상력을 잃는다. 글로벌 기업들이 인도 아웃소싱 업체를 에이전트로 대체하는 흐름은 한국 SI 산업에도 동일한 질문을 던진다. 우리가 수행하는 업무 중 에이전트가 직접 처리할 수 있는 영역은 어디까지인가. 기회의 측면에서, 한국 SI 기업들이 대형 엔터프라이즈 고객과 수십 년간 쌓아온 관계와 시스템 이해도는 에이전트 오케스트레이터 역할로의 전환에서 독보적인 자산이 된다. 글로벌 AI 기업들이 제공하는 범용 에이전트 플랫폼은 특정 산업의 규제 환경, 레거시 시스템의 복잡성, 그리고 고객 조직의 의사결정 구조를 이해하지 못한다. 이 이해도가 한국 SI의 진짜 해자다.

전환의 방향을 구체화하면 세 가지로 수렴된다. **첫째는 에이전트 아키텍처 설계 역량의 내재화다.** 어떤 업무를 에이전트에게 위임하고, 어떤 데이터 접근 권한을 부여하며, 에이전트의 결과를 어떻게 감사하고 검증할지에 대한 설계 역량이 새로운 핵심 서비스 상품이 된다. **둘째는 도메인 특화 에이전트 구축이다. 제조, 금융, 의료, 공공 등 규제와 도메인 지식이 핵심인 영역에서 범용 AI 에이전트가 처리하기 어려운 복잡한 워크플로우를 특화 에이전트로 구현하는 것은 SI의 도메인 전문성이 직접 경쟁력이 되는 영역이다. 셋째는 에이전트 거버넌스 관리다.** 기업 전체의 AI 에이전트가 어떤 시스템에 어떤 권한으로 접속하고 있는지를 중앙에서 관리하는 거버넌스 레이어 구축을 SI가 직접 수행하거나 지원하는 것이다.

기업 리더와 전략기획 담당자들에게 이 글이 전달하는 마지막 질문은 다음과 같다. "우리가 오늘 구독하는 소프트웨어 중, 3년 후에도 동일한 역할로 동일한 비용을 지불할 제품은 무엇인가?" 그리고 역으로, "우리가 오늘 아웃소싱하는 업무 중, 3년 후에도 사람이 수행하고 있을 업무는 무엇인가?" 이 두 질문에 정직하게 답하는 것이, 워크플로우를 가진 자만 살아남는 시대의 전략 수립을 위한 첫 번째 단계다.

# 02 영국의 CTP 지정과 EU의 DMA 게이트키퍼 예비 지정

| 한상기 테크프론티어

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.


유럽에서 전해진 두 가지 소식이 하이퍼클라우드 회사에게 새로운 규제 환경을 만들어내고 있다. 하나는 클라우드 기업이 오히려 환영하고 있고 다른 하나는 앞으로도 많은 논의가 이루어질 사항이다. 먼저 영국에서 발표한 내용을 살펴보도록 한다.

## 영국이 주요 클라우드 기업을 CTP로 지정

영국 정부(재무부, HM Treasury)가 마이크로소프트, 구글 클라우드, 아마존웹서비스, 오라클이라는 특정 법인들을 '핵심 제3자(Critical Third Parties, CTP)'로 지정했고, 이 지정은 7월 13일부터 발효된다고 한다.[6] 감독 대상은 마이크로소프트 아일랜드 사업부, 구글 클라우드 EMEA, 아마존 웹 서비스 EMEA, 오라클 영국 법인 등 4개 법인이며, 금융 부문에 판매하는 시스템적 서비스에만 적용된다.

CTP 감독 체계는 2023년 금융서비스시장법(Financial Services and Markets Act 2023)에 의해 설립되었고, 2025년 1월부터 운영되기 시작했다. 이번에 처음으로 구체적인 지정이 이루어진 것이다. 주목할 점은 이 네 회사가 해당 제도 아래 지정된 최초의 기업들이라는 것이다. 또한 지정 대상이 회사 전체가 아니라 이들이 금융권에 판매하는 시스템적으로 중요한 서비스에만 적용된다는 점도 중요하다.[7]

감독은 단일 기관이 아니라 영란은행(Bank of England), 건전성감독청(PRA), 금융행위감독청(FCA)이 공동으로 담당한다. 지정된 기업들은 복원력 테스트(resilience testing)를 받고, 정기적으로 자체 평가를 실시하며, 주요 사고를 보고해야 하는 의무를 갖는다. 네 개의 기업이 모두 금융기관이 아님에도

6) TradingView, “Microsoft, Google Face UK Cloud Oversight,” Jul 11, 2026
7) TNW, “The UK just declared Microsoft, Google, Amazon and Oracle ‘critical’ to its financial system,” Jul 10, 2026

이런 의무를 부여한 것이다. 각 기관이 갖는 실제 권한을 보면, 영란은행·PRA·FCA는 정보 수집, 복원력 평가, 그리고 필요시 CTP 전용 구속력 있는 규칙의 제정·집행 권한을 갖는다.

이런 지정이 이루어진 핵심 이유는 집중이다. 영국 재무부는 영국 기업의 65% 이상이 인프라 구축을 위해 동일한 4개 공급업체에 의존하고 있다는 사실을 파악했다. 이에 따른 문제의 구조를 다음과 같이 이야기할 수 있다. 한 은행에 국한된 사고는 개별 기관의 실패로 관리할 수 있지만, 여러 은행, 결제회사, 보험사, 거래소가 공유하는 플랫폼 내부에서 발생한 사고는 이들 모두에게 동시에 영향을 미칠 수 있다. 개별적으로는 합리적인 구매 결정들이 누적되어, 어느 단일 고객도 측정할 수 없는 부문 전체의 집중 위험으로 쌓이게 된다는 것이다. 실제로 이 위험이 추상적이지 않다는 걸 보여준 사례가 AWS 장애로 코인베이스를 비롯해 전통 금융기관까지 동시에 마비시킨 사건이다.[8] 클라우드 장애는 이미 은행 앱, 카드 결제 및 거래 도구를 몇 시간 동안 마비시킨 사례가 있다. 작년에 발생한 대규모 AWS 장애로 인해 영국 의원들은 왜 이러한 문제에 대한 조처가 이렇게 오래 걸리는지 의문을 제기하기도 했다.

영국의 이번 조치는 EU를 뒤따른 것이다. EU 감독당국은 2025년 11월에 디지털운영복원력법(DORA) 아래 AWS, 구글 클라우드, 마이크로소프트를 포함한 19개 기술기업을 핵심 제공자로 지정했는데, 두 접근법에는 차이가 있다. EU는 DORA 아래 더 넓은 범위의 19개 기업을 지정한 반면, 영국은 영국 은행·보험 인프라를 떠받치는 4개 하이퍼스케일러에 집중해서 시작했고, 오라클은 EU의 DORA 명단에는 포함되지 않았지만, 영국은 오라클을 추가했다.

DORA(Digital Operational Resilience Act)는 정식으로 EU 규정 2022/2554이며, 2022년 12월 14일에 채택했고 2025년 1월 17일부터 전면 적용했다. 목적은 EU 금융 부문의 디지털 운영 복원력을 강화하는 단일·통합 규제 틀을 만드는 것이다.

핵심 문제의식은 금융기관이 IT 시스템·클라우드·데이터·사이버보안에 전면적으로 의존하게 되면서, 대규모 시스템 장애나 사이버 공격이 곧바로 금융 안정성 위협으로 번질 수 있게 되었다는 것이다. DORA는 이 디지털 리스크를 자본과 유동성 리스크처럼 정식 규제 대상으로 끌어올린 법이다. DORA는 은행, 보험사, 투자회사부터 암호자산 서비스 제공자(CASP)까지 폭넓은 범위의 금융기관에 적용한다. 중요한 특징은 두 층위에 동시에 의무를 부여하는데, 하나는 금융기관 자체에게 직접 부여하는 의무이며, 다른 하나는 핵심 ICT 제3자 제공자(CTPP)에 대한 EU 차원의 감독 체계이다.

8) Coinbase, “A postmortem of our May 7, 2026 outage,” Jun 1, 2026

2025년 11월 18일, 유럽감독당국(ESAs), 즉 유럽은행감독청(EBA), 유럽보험·직역연금감독청(EIOPA), 유럽증권시장감독청(ESMA)이 DORA 제31조 9항에 따라 핵심 ICT 제3자 제공자(CTPP) 최초 명단을 발표했다. 총 19개 제공자가 지정되어 EU 차원의 직접 감독 대상이 되었는데, 명단에는 하이퍼스케일 클라우드 제공자, 데이터센터, 인프라·네트워크 제공자, 금융서비스 전용 기술 제공자를 포함했다. 이 명단은 영구적이지 않고 ESAs가 매년 명단을 갱신해 발표한다. 영국의 CTP 제도는 DORA와 문제의식·설계가 매우 유사하며, 영국이 EU를 뒤따른 성격을 갖고 있다.

그러나 오해하지 말아야 하는 점은 규제 당국이 클라우드 제공사를 직접 감독한다고 해서, 금융기관 자신의 책임이 사라지는 것은 아니다. 제공사 수준의 직접 개입이, 고객사의 아키텍처, 접근 통제, 연속성 계획, 복구 실행을 다른 누군가의 책임으로 만들어주지는 않는다는 것이다. 오히려 가장 위험한 해석은 은행들이 이제 규제 당국이 자사의 클라우드 공급사를 검증해준다고 믿는 것이라는 지적이 있다.

주의를 끄는 점은 CTP로 지정된 회사들이 저항하지 않는다는 점이다. 구글 클라우드는 이 프레임워크가 잘 이행된다면 영국 금융 생태계의 장기적 복원력을 강화하고 당사자 간 이해, 투명성, 신뢰를 높일 수 있다고 밝혔고, 마이크로소프트도 관련 감독 요건 준수를 약속했다. 규제된 금융 영역에 더 깊이 판매하려는 기업 입장에서는 '핵심 인프라'로 인증받는 것이 싸워서 물리쳐야 할 부담이 아니라 갖고 싶은 라벨이기 때문이다. 즉 CTP 지정은 진입 장벽이자 일종의 해자(moat) 역할을 할 수 있을 것으로 본다.

이 조치는 "빅테크 클라우드 4사가 이제 은행처럼 금융 안정성 관점에서 규제받는 인프라가 되었다"는 의미로 해석해야 한다. 영국이 클라우드를 핵심 인프라로 규제하기 시작했다는 점에서 규제 경계선이 애플리케이션 계층에서 클라우드 플랫폼 자체로 이동하고 있다는 상징적 전환점이 된 것으로 봐야 한다.

더 근본적인 메시지는 권력에 관한 것이다. 규제 당국은 10년 동안 은행들이 너무 커서 망할 수 없다고 우려해 왔다. 이제는 은행들이 사용하는 클라우드 회사들이 너무 중심적인 위치를 차지하게 되어 더 이상 무시할 수 없게 되었다는 점이 우려거리이다. 7월 13일부터 영국은 클라우드 기업을 그런 시각으로 전체 금융 시장을 바라볼 것이다.

## EU 집행위(EC)가 AWS와 애저를 게이트키퍼로 지정

이 소식은 6월 말에 나온 것으로 앞에서 설명한 영국과는 다른 각도의 규제를 제시한 것이다. 유럽집행위(EC)는 2026년 6월 25일, 아마존의 AWS와 마이크로소프트의 애저를 디지털시장법(DMA)상 '게이트키퍼(gatekeeper)'로 지정해야 한다는 예비 입장을 두 회사에 통보했다.[9] 집행위는 AWS와

애저가 각각 EU 최대와 2위 클라우드 서비스로서, EU 내 기업과 고객을 잇는 중요한 관문이라고 예비 판단을 내렸다는 것이다.

여기서 가장 중요한 대목은 EC 스스로 이들이 DMA의 정량적 지정 기준을 충족하지 못한다는 점을 인정하면서도 지정을 추진한다는 것이다. 원래 DMA는 기계적으로 설계된 법이다. 일정 시가총액, EU 내 월간 이용자 수 같은 정량 기준을 넘고 '핵심 플랫폼 서비스'를 운영하면 자동으로 게이트키퍼로 추정되거나 지정되는 방식이다. 애플, 구글, 메타, 아마존 마켓플레이스, 마이크로소프트, 바이트댄스, 부킹닷컴이 이 방식으로 편입됐다.

그런데 이번에는 AWS나 애저가 정량 기준을 충족하지 못했는데도 EC가 게이트키퍼 지위를 적용했으며, 이는 이 규제 역사상 최초이다. 대신 EC는 정성적 평가에 의존했는데, 그 근거는 관문 기능, 고착화된 시장 지위, 그리고 특히 락인 효과와 높은 전환비용이다. 두 회사 모두 방대하고 고착된 이용자 기반을 갖추고 있고, 대규모 생태계와 더불어 락인 효과·높은 전환 비용의 혜택을 누리는 것으로 보인다는 것이다.[10] 이번에 집행위가 제시한 세 가지 기준은 EU 역내시장에 대한 중대한 영향, 비즈니스 이용자가 최종 이용자에 도달하는 중요한 관문 기능, 고착적·지속적 시장 지위로 구조적 판단이라고 볼 수 있다.

이 사안에서 주목할 만한 부분은, EC가 AI를 클라우드 지배력을 굳히는 핵심 요소로 명시했다는 것이다. EC는 AI 도구와 AI-클라우드 파트너십을 클라우드 조달의 결정적 요인으로 지목하면서, 인프라 락인을 AI 스택과 직접 연결지었다. 부위원장 헨나 비르쿠넨(Henna Virkkunen)은 클라우드 서비스를 AI의 전제조건이라고 표현했다고 한다. 즉 이번 조치는 단순한 클라우드 경쟁 문제가 아니라, EU가 AI를 떠받치는 계층 자체를 규제 대상으로 끌어들인 사건이다.

실제로 여러 매체가 이를 소비자 플랫폼(검색·SNS·앱스토어)에 머물던 DMA가 처음으로 클라우드 인프라, 즉 AI를 지탱하는 층으로 확장된 것이라고 평가하고 있다. 그동안 이 이슈에 진행한 절차와 앞으로 이루어질 추가 액션에 대한 시간표를 보면 다음과 같다.

- 집행위가 2025년 11월 18일에 두 건의 시장조사(market investigation)를 개시했다.
- 약 7개월의 조사 후 2026년 6월 25일, 예비 입장 발표

9) EC, “Commission reaches preliminary position that Amazon's and Microsoft's market leading cloud services should be designated under the DMA,” Jun 25, 2026
10) The Register, “European Commission lines up Amazon and Microsoft for cloud gatekeeper status,” Jun 25, 2026

- AWS와 애저는 2026년 9월까지 서면 의견을 제출하고 청문을 요청할 수 있다.
- DMA 시장조사는 개시 후 12개월 내에 종결해야 하므로, 집해위는 2026년 11월까지 최종 결정을 내려야 한다. 실제 결정은 10월말로 예상한다.
- 최종 지정되면 게이트키퍼는 6개월 내에 의무를 이행해야 한다.

게이트키퍼 지정이 확정되면 상호 운용성, 데이터 이동성 보장, 자사 서비스 우대 금지, 이그레스 수수료 (데이터 반출 비용) 제한 등이 부과되며 이를 위반시 전 세계 매출의 최대 10%까지 과징금이 가능하다. 이 의무는 기업이 고객을 묶어 두기 위한 매커니즘에 적극 대응하는 것이기 때문에 기업들에게 매우 아픈 지점이다.

이 조치에 대해 아마존은 2022년 이후 약 15% 점유율을 가진 200개 이상의 EU 클라우드 경쟁사가 존재한다는 의뢰 연구를 근거로 반박하고 있고, 마이크로소프트는 구글 클라우드의 AI 기반 성장을 지목하고 있다. 게이트키퍼 사실 자체를 다투기보다 시장 획정의 오류를 지적하는 전략을 취하고 있다.

더 큰 그림으로 보면 이 DMA 절차는 EU의 광범위한 '소버린 클라우드' 드라이브와 나란히 진행된다. 집행위는 2026년 초 최대 1억 8천만 유로 규모의 6년 클라우드 계약을 STACKIT, 스케일웨이(Scaleway), 포스트 텔레콤, 프록시머스(Proximus) 등 유럽 업체에만 배타적으로 발주했고, 프랑스는 국가 보건데이터허브를 애저에서 스케일웨이로 이전했다. 또한 이 사안은 EU 클라우드와 AI개발법(EU Cloud and AI Development Act)과 데이터법의 전환·이그레스 조항과도 맞물려, 모두 클라우드 락인이라는 동일한 표적을 향하고 있는 것이다.

## 마무리하며

같은 두 회사를 놓고 영국은 훨씬 가벼운 길을 택했다. 영국 CMA는 2026년 5월 14일 마이크로소프트의 비즈니스 소프트웨어 생태계에 대해 전략적시장지위(SMS) 조사를 개시했지만, AWS는 영국에서 SMS 조사를 받지 않았다. 결과적으로 양쪽 시장에서 클라우드를 구매하는 기업은 이제 집행 가능한 의무를 런던이 아니라 EU에서 찾아야 한다.

AWS와 애저가 게이트키퍼로 지정되면, 두 회사는 데이터 이동성·상호운용성 보장 및 자사 우대 금지 의무를 지게 되고, EU 집행위가 이를 실행하는 가에 대한 감독하고 위반시 과징금을 부여하는 공적 집행의 기관이다. EU 고객은 그 결과로 이러한 보호를 누리며, 필요시 회원국 법원에서 사적 집행을 통해 이를 다툴 수도 있다. 영국 고객은 제공사가 자발적으로 제시한 계약상 약속에 의존해야 한다. 두 규제 축을 EU DORA와 함께 정리하면 다음 그림과 같다.

| 구분 | 규제 논리 | 근거 법 | 관할 | 이번 사안 |
|---|---|---|---|---|
| 영국 CTP | 운영 복원력·금융안정 | FSMA 2023 | 영란은행 ·PRA·FCA | MS·구글·AWS·오라클 지정 (7/13 발효) |
| EU DORA | 운영 복원력·금융안정 | DORA | 유럽 금융감독당국 | 19개사 지정 (2025/11) |
| **EU DMA** | **경쟁·시장지배력** | **DMA** | **유럽위원회** | **AWS·Azure 게이트키퍼 예비지정 (6/25)** |

그림 2 두 규제 방안과 EU DORA와 비교 표

CTP와 DORA는 클라우드가 무너지면 금융시스템이 위험하다는 복원력과 안정성에 대한 논리이고, DMA 건은 소수 클라우드가 시장을 장악한다는 경쟁과 반독점의 논리이다. 대상 기업은 겹치지만 규제가 노리는 위험의 성격이 전혀 다르지만, 모두에서 공통적으로 등장하는 배경이 집중과 AI 의존이라는 점이, 지금 클라우드-AI 인프라 규제의 시대적 특징이라고 볼 수 있다.

# 03 토큰 이코노미시대의 AI 핀옵스 전략

| 윤대균 아주대학교

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.


## 1. 들어가며

추론 단가는 급격히 떨어지고 있지만 기업의 AI 비용은 오히려 급증하고 있다. 생성형 AI가 POC단계를 지나 이제 본격적인 활용 단계로 넘어가면서, AI 모델을 실제 업무에 사용하기 위한 “추론”이 상시 업무에 필요한 대표적인 워크로드로 자리 잡았다. 그 결과 일부 기업은 AI 사용료로 월 수천만 달러 규모의 청구서를 받기 시작했다. 토큰당 단가가 내려가면 총비용도 내려가리라는 것이 상식이지만, 현실은 반대로 움직인다. 단가 하락 속도보다 사용량 증가 속도가 더 빠르기 때문이다. 자원의 사용 효율이 높아질수록 오히려 총소비가 늘어난다는 제번스 패러독스(Jevons paradox)가 AI 추론에서 그대로 재현되고 있다.

동일한 성능을 기준으로 보면 추론 단가는 연 수십 배 속도로 하락하고 있다. AI 연구기관 에포크 AI(Epoch AI)는 6개 벤치마크에서 특정 성능 수준에 도달하는 데 필요한 최저 단가가 얼마나 빨리 떨어졌는지를 측정했다. 그 결과 하락 속도는 성능 구간에 따라 연 9배에서 900배까지 편차가 컸고, 중앙값은 연 50배였다. 예를 들어 박사급 과학 문제(GPQA Diamond)에서 GPT-4 수준의 성능에 도달하는 단가는 연 40분의1로 떨어졌다.[11)]

여기서 말하는 성능 구간이란 특정 난이도의 과제를 특정 품질 이상으로 풀어내는 능력 수준을 뜻한다. 벤치마크마다 목표 점수를 정해 두고, 그 점수 이상을 내는 가장 저렴한 모델의 단가를 추적/분석하는 방식이다. 낮은 성능 구간은 요약이나 분류처럼 비교적 쉬운 과제에 해당하며, 이미 값싼 소형 모델로도 충분히 도달할 수 있어 단가가 이미 바닥에 가깝다. 반면 높은 성능 구간은 박사급 과학이나 고급 수학처럼 어려운 과제에 해당하며, 최근까지도 고가의 최신 모델이라야 도달할 수 있었으나 그만큼 하락

11) Ben Cottier et al., “LLM inference prices have fallen rapidly but unequally across tasks”, Epoch AI, Mar. 2025

여지가 커서 지금도 빠르게 값이 떨어지고 있다.

추론 비용을 이해하는 기본 단위는 토큰이다. 토큰은 모델이 처리하는 언어의 최소 조각으로, 입력 토큰과 출력 토큰으로 나뉜다. 상용 대규모 언어모델(LLM)의 가격은 통상 100만 토큰당 비용(CPM)으로 표시된다. 서로 다른 하드웨어와 처리량을 하나의 비교 가능한 숫자로 환산해 주기 때문에, CPM은 추론 경제를 가늠하는 표준 지표로 자리 잡았다.

## 2. 토큰 이코노미와 추론 비용의 역설

토큰 이코노미란 AI 서비스의 소비량과 원가가 토큰으로 계량되는 경제 구조를 뜻한다. 전력 사용량이 킬로와트시로 계량되고 그만큼 요금이 매겨지듯, AI에서는 입력과 출력이 모두 토큰으로 계량되고 토큰 수에 단가를 곱해 요금이 매겨진다. 사용자가 던진 프롬프트는 입력 토큰으로, 모델이 내놓은 응답은 출력 토큰으로 환산되며, 둘의 합에 단가를 곱한 값이 청구액이 된다. 따라서 어떤 기능이 소비하는 토큰량은 곧 그 기능의 원가이며, 서비스 기획자와 경영자는 토큰 소비를 매출과 이익 구조에 연동해 관리해야 한다.

토큰 이코노미에는 두 가지 구조적 특성이 있다. 첫째, 모델이 한 번에 처리할 수 있는 토큰 수인 컨텍스트 창이 유한하다. 컨텍스트는 무한히 늘릴 수 있는 공간이 아니라 값이 매겨진 희소 자원이다. 창을 채울수록 입력 토큰 비용이 늘고, 어텐션 연산량은 토큰 수의 제곱에 비례해 커지며, 너무 길어지면 중요한 정보가 묻혀 오히려 응답 품질이 떨어진다. 따라서 컨텍스트에 무엇을 담을지 자체가 비용과 품질을 좌우하는 설계 문제가 된다.

둘째, 같은 의미에도 불구하고 언어에 따라 토큰 수가 다른 다국어 토큰화 격차가 존재한다. 한국어는 같은 내용을 표현하는 데 영어보다 많은 토큰을 소비한다. 측정에 따르면 한국어는 동일한 정보를 담는 데 영어의 약 2.4배에 이르는 토큰을 쓰며, 이 때문에 토큰당 과금 서비스에서 한국어 사용자는 영어 대비 1.5배에서 2배가량 비용을 더 부담한다.[12] 조사와 어미가 겹겹이 붙는 한국어의 교착어 특성이 영어 중심 토크나이저에는 최적화되어있지 않기 때문이다.

이 구조에서 단가가 내려가도 총비용이 늘어나는 이유는 사용량이 단가 하락보다 훨씬 빠르게 증가하기 때문이다. 이것이 추론 경제의 핵심 역설이다. 새로운 성능 구간이 저렴해질 때마다 그 가격에서 비로소

12) Anthony Shaw, “Working with Chinese, Japanese, and Korean text in Generative AI pipelines”, Mar 12, 2024

가격대비 쓸만한 새로운 활용 사례가 열리고, 그 사례들이 다시 토큰 소비를 늘린다. 가격이 내려간 만큼, 혹은 그 이상으로 수요가 팽창하는 구조다.

기업의 실제 지출 데이터가 이를 뒷받침한다. 멘로 벤처스(Menlo Ventures)의 조사에 따르면 기업의 생성형 AI 지출은 2023년 17억 달러에서 2024년 115억 달러, 2025년 370억 달러로 늘었다. 한 해 만에 3.2배, 2년 만에 약 22배로 커진 것이다.[13] 토큰당 단가가 90% 넘게 떨어지는 동안 총지출은 오히려 수십 배로 불어났다는 뜻이다. 전체 시장으로 보면 가트너는 2026년 전 세계 AI 지출이 전년 대비 44% 늘어난 2조 5,200억 달러에 이를 것으로 전망한다.[14] 그림 3은 단가와 총비용이 반대 방향으로 움직이는 이 역설을 요약한 것이다.


**그림 3** 추론 비용의 역설

수요 폭증을 이끄는 주요 요인은 추론특화 모델과 에이전트다. 여기서 "추론특화 모델"이라 함은 답을 내기 전에 내부적으로 단계적 사고 과정을 길게 생성하는 모델을 말한다. 답을 내기 전에 문제를 쪼개고 중간 계산을 거치며 스스로 점검하는 과정을 밟는다. 그 덕에 수학이나 코딩, 논리 추론처럼 어려운 과제에서 정확도가 크게 높아진다. 대신 그 중간 사고 과정에서 토큰을 대량으로 소비하기 때문에 비용이 훨씬 커진다. 요컨대 간단한 요약이나 문의 응대는 즉답형 모델로 충분하지만, 복잡한 분석은 추론특화 모델이 유리하며, 그 대가는 대량의 토큰 소비이다.

에이전트는 여기서 한 걸음 더 나아간다. 자율적으로 작업을 분해하고 도구를 호출하며 결과를 검증하고 스스로 교정하는 과정에서, 하나의 사용자 요청이 수십 번의 모델 호출로 확장되기도 한다. 에이전트형

13) Menlo Ventures, "2025: The State of Generative AI in the Enterprise", Dec. 09, 2025.
14) Gartner, "Gartner Says Worldwide AI Spending Will Total $2.5 Trillion in 2026", Jan. 15, 2026

워크로드가 일반 챗봇 대비 작업당 5배에서 25배 더 많은 비용을 소요한다는 분석도 나와 있다.[15] 그림 4에서 볼 수 있듯이 과거의 챗봇 스타일에서는 질문 1건에 모델 호출 한 번으로 짧게 답했다면, 에이전트형 처리 방식에서는 질문을 여러 하위 작업으로 나누고, 사내 문서를 검색하며, 답의 정확성을 스스로 검증하고, 필요하면 재작성한다. 이를 위해 내부에서는 열 번 이상의 모델 호출과 수만 개의 토큰이 소모된다. 응답 품질은 분명히 높아지지만 질문 건당 원가도 그만큼 커진다는 뜻이다. 결국 에이전트 워크플로우를 어떻게 설계하고 통제하느냐에 따라 총비용이 결정되며, 이를 최적화하기 위한 전략적 접근이 필요하다. 이를 AI 핀옵스 관점에서 살펴보자.

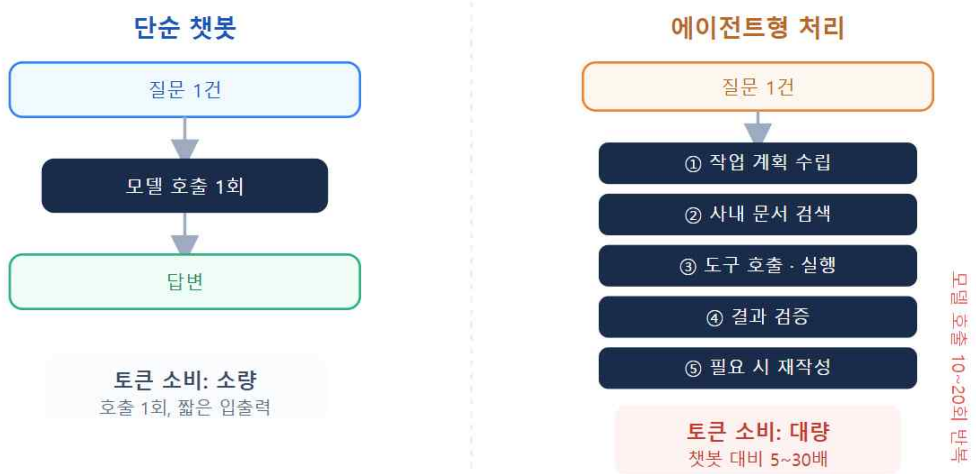


그림 4 같은 질문, 다른 토큰 소비

## 3. 추론 경제의 운영: AI 핀옵스

추론 비용의 최적화는 설계 단계에서부터 적절한 거버넌스하에 이루어져야 한다. 우선 비용을 측정하는 기준부터 바꿔야 한다. 토큰당 단가는 물론 필요하지만 이것만으로는 충분하지 않다. 최선의 결과를 추구한다는 명목으로 최신 모델을 기본값으로 두고, 추론 강도를 최고로 설정하거나, 또는 문맥 창을 크게 잡는 것은 모두 값비싼 청구서로 귀결된다. 올바른 목표는 토큰당 비용이 아니라 "필요한 품질과 지연, 위험 수준에서 성공적 결과 하나를 얻는 데 드는 비용"이어야 한다.[16]

비용 최적화를 위해서는 비용 구조를 워크플로우, 모델, 런타임, 인프라의 네 계층으로 구분하여 각 계층별 거버넌스, 즉 핀옵스로 전 과정을 상시 점검해야 한다(그림 5). 워크플로우 계층에서 불필요한

15) TechAhead, “The Inference Cost Trap: Why Your AI Agent Economics Break At Scale”, Jun 8, 2026
16) Spheron, “AI Inference Cost Economics in 2026: GPU FinOps Playbook”, Apr 4, 2026.

호출을 걷어내고, 모델 계층에서 적정 모델을 고르며, 런타임 계층에서 실행 효율을 높이고, 인프라 계층에서 계산 자원을 알맞게 배치하는 구조다. 상위 계층일수록 영향을 받는 대상이 크므로 절감 폭도 크다. 다만 상위 계층일수록 재설계 난이도와 그에 따른 품질 변화 위험이 크기 때문에 거버넌스 실행시 적용 순서는 성과 대비 노력을 함께 고려해야 할 것이다.

각 계층에 대해 조금 더 들여다보겠다.

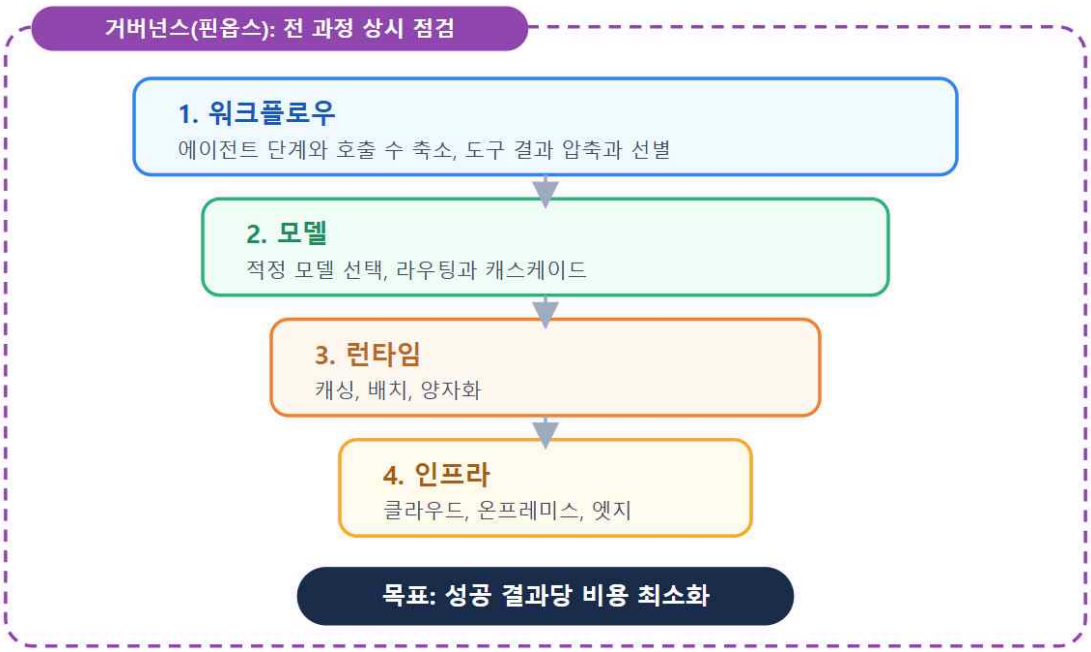


그림 5 추론 비용 최적화의 4계층

### 3.1 워크플로우 계층: 에이전트와 도구 호출 최적화

워크플로우 계층은 가장 상위이자 레버리지가 가장 큰 최적화 지점이다. 거버넌스의 출발점은 과제에 정말 에이전트가 필요한지 되묻는 것이다. 단순 조회나 요약같은 작업은 자율 반복 없이 단일 호출로 처리하고, 에이전트를 쓰더라도 불필요한 반복 루프를 없애고 조기 종료 조건을 두어 호출 수와 단계를 줄인다.

도구 호출이 만들어 내는 컨텍스트 팽창을 관리하는 것도 핵심이다. MCP처럼 외부 도구를 붙이면, 데이터베이스 조회나 검색 결과가 통째로 컨텍스트에 쌓여 토큰 소비가 급증한다. 도구가 돌려준 결과에서 의미 있는 필드만 추리거나 요약해 다음 단계로 넘기고, 필요할 때만 상세 정보를 다시

불러오는 방식의 설계가 필요하다. 특히 긴 대화나 멀티 에이전트 구조에서는 컨텍스트 관리 자체가 최적화의 핵심이다. 앤스로픽은 에이전트 협업 시 '컨텍스트 엔지니어링'의 중요성을 강조한다. 예컨대 전체 맥락을 통째로 주고받는 대신 하위 작업을 서브 에이전트에 위임하고, 에이전트 간에는 핵심 요약만 전달하며, 오래된 맥락은 자동 요약(Compaction)하여 토큰 낭비를 막는 식이다.[17)]

최근 많은 관심을 받고 있는 하네스 엔지니어링도 워크플로우 최적화 관점에서 재조명해 볼 필요가 있다. 하네스는 실행 루프와 도구 호출, 컨텍스트 관리, 메모리, 가드레일을 묶어 모델을 실제 원하는 방향으로 작동하는 에이전트로 만드는 역할을 한다. 하네스 엔지니어링은 주로 결과의 품질과 신뢰성을 높이려는 흐름에서 각광받고 있지만, 토큰 이코노미 관점의 워크플로우 최적화와 사실상 같은 설계 표면에 있다. 하네스를 어떻게 설계하느냐가 곧 토큰 경제를 좌우한다고 보아도 과언이 아니다. 검증 루프나 다중 후보 생성처럼 품질을 위해 토큰을 더 쓸 경우에 이에 대한 정당성을 확보하는 것도 어떻게 보면 설계 단계에서 거버넌스 행위에 해당한다고 볼 수 있다. 즉 훌륭한 하네스 엔지니어링은 토큰을 사후 정산해야 할 '소모 비용'이 아니라, 시스템 설계 초기부터 최적화하고 통제해야 할 '핵심 자본'으로 다루는 것이다.

### 3.2 모델 계층: 적정 모델 선택과 라우팅

모델 계층의 원칙은 모든 요청에 최고 성능 모델을 쓰지 않는 것이다. 과제 난이도에 맞는 가장 값싼 모델을 고르는 것이 출발점이다. 이를 자동화한 기술이 모델 라우팅과 캐스케이드(cascade)다. 라우팅은 들어온 질의를 그것을 제대로 처리할 수 있는 가장 저렴한 모델로 보내는 방식이고, 캐스케이드는 우선 소형 모델이 처리해 보고 확신이 낮을 때만 대형 모델로 승급시키는 방식이다. 승급 여부는 소형 모델이 내놓은 답의 불확실성, 즉 토큰 확률 분포가 얼마나 한쪽에 몰려 있는지를 근거로 판단한다.

적절한 라우팅을 통해 최고 모델 품질의 95%를 유지하면서 질의의 85%를 값싼 모델로 처리해, 워크로드에 따라 45%에서 85%의 비용을 줄일 수 있다는 분석이 있다. 실제 운영 환경에서도 40%에서 70%의 절감이 흔히 보고된다.[18)] 더 나아가 모든 것을 GPU로 돌릴 필요는 없으며, 실제 가치를 내는 상당수 AI 워크로드는 범용 CPU로도 충분히 돌아간다고 지적하기도 한다. 과제 수준이나 규모 대비 과잉 자원을 걷어내는 것만으로도 큰 절감이 가능하다는 뜻이다.

17) Anthropic, “Effective context engineering for AI agents”, Sep 29, 2025
18) LeanLM, “LLM Model Routing: Automatically Send Every Query to the Cheapest Capable Model”, May, 2026

### 3.3 런타임 계층: 캐싱과 서빙 최적화

런타임 계층에서는 같은 모델이라도 실행 방식을 바꿔 원가를 낮출 수 있다. 대표적으로 프롬프트 캐싱을 들 수 있다. 여러 요청에 공통으로 들어가는 시스템 프롬프트나 긴 문맥을 캐시에 저장해 두면, 매번 같은 입력을 새로 처리하지 않아도 된다. 앤스로픽은 캐시에서 읽는 입력 토큰에 90% 할인을 적용한다. 한 보안 기업은 캐시 적중률을 7%에서 84%로 끌어올려 전체 LLM 비용을 59%에서 70%까지 줄였는데, 이때 캐시에서 처리된 토큰만 98억 개에 이르렀다고 한다.[19)]

서빙 엔진 차원의 최적화도 효과가 크다. 연속 배치(continuous batching)는 먼저 끝난 요청 자리에 새 요청을 곧바로 채워 GPU를 놀리지 않게 하고, 양자화는 모델 가중치를 FP8이나 INT8 같은 저정밀 형식으로 바꿔 메모리와 연산을 아낀다. 추론적 디코딩(speculative decoding)은 작고 빠른 초안 모델이 다음 토큰을 미리 추측하고, 본 모델이 한 번에 검증하게 해 생성 속도를 높인다. 이런 방식으로 처리량이 많아지면 같은 하드웨어로 더 많은 토큰을 처리하므로 토큰당 원가가 내려간다.

### 3.4 인프라 계층: 클라우드와 온프레미스의 조합

인프라 계층에서는 워크로드 특성에 맞는 연산 자원을 선택하는 것이 중요하다. 변동성이 크거나 실험 성격이 강한 워크로드, 최신 모델을 빠르게 써야 하는 경우에는 퍼블릭 클라우드가 유리하다. 반면 대량으로 상시 반복되는 운영 추론은 온프레미스, 저지연이 중요한 처리는 엣지가 적합하다. 이에 많은 기업이 하나의 플랫폼에 몰아넣는 대신 워크로드마다 최적 플랫폼을 조합하는 하이브리드 컴퓨트 포트폴리오로 옮겨가고 있다.

최적의 하이브리드 조합을 위한 자동화 도구도 점차 확대되는 추세다. 이런 도구가 확산되면서 인프라 조달은 주기적이고 수작업에 의존하던 방식에서 상시적이고 알고리듬에 기반한 방식으로 이동하고 있다. 워크로드 수요와 비용 변동, 성능 요구를 실시간으로 반영해 모델과 인스턴스를 동적으로 배분하는 단계로 나아가는 것이다. 이는 기존 클라우드 네이티브 컴퓨팅 패러다임이 AI 연산 최적화를 지원하는 형태로 빠르게 진화하고 있음을 시사한다.

### 3.5 거버넌스 계층: 핀옵스와 아키텍처 규율

앞의 네 계층이 제대로 작동하려면 비용 가시성이 전제되어야 한다. 어떤 팀과 기능, 어떤 모델이 토큰을 얼마나 소비하는지 분해해 볼 수 없으면 최적화 대상을 특정할 수 없다. 클라우드 인프라 지출을

19) ProjectDiscovery, “How We Cut LLM Costs by 59% With Prompt Caching”, Apr 10, 2026

부서나 프로젝트 단위로 할당하고 정산해 온 전통적 핀옵스처럼 AI 워크로드도 호출 단위로 소비 토큰과 비용을 기록하고 서비스와 기능별로 귀속시키는 측정 체계가 먼저 갖춰져야 한다. 학습 비용과는 달리 추론 비용은 예측이 어렵고 빠르게 변화하므로 주 단위 혹은 그보다 작은 단위의 지속적 점검으로 이상 급증을 조기에 잡아야 한다.

거버넌스는 조직 구조로도 내재화 되어야 한다. 아키텍처 리뷰 보드를 두어 신규 AI 프로젝트가 비용과 성능, 거버넌스, 위험 요소 등의 판단 기준에 따라 적정 인프라 위에서 돌아가는지 심사한다. 어떤 과제는 외부 API 호출로, 어떤 과제는 온프레미스 자원으로 처리하도록 초기부터 규율하는 것이다. 이런 흐름을 따라가기 위해서는 조직의 구성원이 추론 경제와 하이브리드 원가 구조를 이해할 수 있는 역량을 갖추어야 한다.

## 4. 시사점

추론 경제에서의 절대 금기 사항은 단가 하락이 비용 절감으로 이어질 것이라는 막연한 기대감이다. 토큰당 가격이 떨어질 것을 전제로 AI 예산을 짜는 것은 매우 위험한 시나리오다. 사용량이 단가보다 빠르게 늘어나는 한 총비용은 계속 증가한다. 값싼 추론은 더 많은 활용을 부르고, 더 많은 활용은 더 큰 청구서로 돌아온다.

첫째 과제는 토큰을 원가의 기본 지표로 삼는 운영 체계를 구축하는 것이다. 서비스 기획부터 운영까지 모든 단계에서 소비하는 토큰량을 원가로 인식하고, 결과당 비용을 표준 성과지표로 삼아야 한다. 둘째 과제는 비용 거버넌스를 설계 단계에 내재화하는 것이다. 어떤 모델을 어떤 인프라에서 어떤 강도로 돌릴지를 초기부터 규율하는 아키텍처 심사 체계가 필요하다. 셋째 과제는 컴퓨트 배치의 유연성 확보다. 모든 워크로드를 클라우드나 GPU에 몰아넣는 대신 클라우드와 온프레미스, 엣지를 워크로드 특성에 맞게 조합하는 하이브리드 전략이 원가 경쟁력을 좌우한다.

국내 기업과 공공부문에도 시사하는 바가 크다. 생성형 AI활용을 본격적으로 확산하는 국면에서 추론 비용을 체계적으로 관리하는 운영 역량은 우선순위에서 다소 밀려나 있다. 도입 초기부터 토큰 기반 원가 관리와 핀옵스 체계를 함께 세운다면, 뒤늦게 폭증하는 비용에 대응함으로써 맞닥뜨릴 문제를 최소화할 수 있다. 특히 데이터 주권이나 규제 대응이 중요한 공공과 금융 영역에서는 온프레미스와 소버린 클라우드가 비용과 통제를 동시에 잡는 현실적 선택지가 될 수 있다. 추론 단가가 저렴해질수록 이를 얼마나 규율 있게 쓰느냐가 조직의 AI 경쟁력을 가르게 될 것이다.

# 04 2026년 클라우드 솔루션 리포트: IAM/ID 플랫폼

| 정채상 KAIST

본 글은 한국지능정보사회진흥원의 지원을 받아 작성되었습니다.


## 들어가며: 그 에이전트는 누구인가?

지난달 글에서는 "AI 에이전트가 API의 새로운 소비자가 되었다"는 진단에서 출발했다. 에이전트가 사람을 대신해 도구를 호출하고 사내 데이터에 닿기 시작하면서, API 게이트웨이가 그 트래픽이 반드시 통과하는 통제점으로 돌아왔다는 이야기였다. 게이트웨이가 "어느 경로로 들어오는가"를 강제한다면, 그 앞에는 반드시 해야 할 다른 질문이 남는다. 지금 이 API를 호출하는 에이전트는 대체 누구인가? 어떤 권한으로 움직이며, 그 권한은 누가 부여했는가?

전통적인 신원 및 접근 관리(IAM, Identity and Access Management)는 이 질문에 답할 준비가 되어 있지 않은데, 이는 지난 20년간 IAM이 관리해 온 대상은 사람이었기 때문이다. 싱글 사인온(SSO), 다중 인증(MFA), 패스워드 정책은 모두 "인간이 시스템에 로그인한다"는 전제 위에서 구현되어 있다. 그러나 2026년 엔터프라이즈 환경에서 실제로 시스템에 접근하는 주체의 대다수는 사람이 아닌데, 서비스 계정, API 키, OAuth 토큰, 인증서, 그리고 이제 AI 에이전트까지 — 비인간 ID(NHI, Non-Human Identity)가 인간 신원보다 수십에서 수백 배 많다.

사이버아크(CyberArk)를 인수한 팔로알토 네트웍스가 2026년 5월 차세대 플랫폼 이디라(Idira)를 출시하며 인용한 수치는 "머신과 AI 신원이 인간을 109 대 1로 초과한다"였다.[20] 비인간 ID를 전문으로 추적하는 NHI 관리 그룹(NHI Management Group)의 2026년 보고서는 이 비율을 조직에 따라 45 대 1에서 최대 100 대 1로 잡았고, 머신 신원의 약 70%가 문서화되지 않았거나 수개월간 쓰이지 않은 채로도 여전히 높은 권한을 쥐고 있다고 정리했다.[21] 인간의 수십 배에 이르는 신원이,

20) Palo Alto Networks. "Palo Alto Networks Introduces Idira, the Next-Generation Identity Security Platform Built for the AI Enterprise." Palo Alto Networks Press, 2026.05.12. https://www.paloaltonetworks.com/company/press/2026/palo-alto-networks-introduces-idira--the-next-generation-identity-security-platform-built-for-the-ai-enterprise

대부분 과도한 권한을 쥔 채, 아무도 들여다보지 않는 곳에서 살아 움직이는 것이다.

공격자는 이 사각지대를 이미 알고 있으며, 2026년 4월 가트너 IAM 서밋은 이를 간결하게 경고했다. "공격자는 더 이상 뚫고 들어오지 않는다. 그들은 로그인한다".[22] 소포스(Sophos)가 17개국 5,000명을 조사한 2026년 보고서에서 71%의 조직이 지난 1년간 하나 이상의 신원 관련 침해를 겪었고, 그중 41%의 침해에서 비인간 ID 관리 부실이 원인으로 지목되었다. 침해 한 건의 평균 복구 비용은 164만 달러에 달했고[23], 시크릿 스프롤(Secrets Sprawl, 자격 증명이 코드와 설정 곳곳에 흩어져 통제 불능이 되는 현상)을 추적하는 깃가디언(GitGuardian)은 2025년 한 해에만 공개 깃허브 커밋에서 2,865만 개의 하드코딩 된 시크릿이 새로 노출됐고, 그중 AI 서비스 관련 유출은 81% 급증했다고 보고했다.[24]

시장은 이 변화를 대형 보안 벤더들의 큰 인수합병으로 보여준다. 2026년 상반기에 이어진 가장 큰 인수들이 모두 같은 방향을 가리켰는데, 크라우드스트라이크는 1월 실시간 동적 인가 기술을 가진 SGNL을 인수한다고 발표했고[25], 팔로알토 네트웍스는 2월 사이버아크 인수를 완료했으며[26], 시스코는 5월 비인간 ID 발견 전문 기업 아스트릭스(Astrix)의 인수 의향을 밝혔다.[27] 이는 방화벽과 엔드포인트에 집중하던 보안의 축이 신원으로 옮겨 가고 있다는 신호로 볼 수 있다. 제로 트러스트가 오래전부터 "신원이 새로운 경계선"이라고 말해 왔다면, 2026년의 인수들은 그 내용을 자본의 흐름으로 뒷받침하는 셈이다.

---

21) NHI Management Group. "The Non-Human Identity Security Perimeter 2026." nhimg.org, 2026. https://nhimg.org/nhi-101/non-human-identity-security-perimeter-2026
22) GitGuardian. "Gartner IAM Summit 2026: Identity Expanded Faster Than Most Programs Did." GitGuardian Blog, 2026.04. https://blog.gitguardian.com/gartner-iam-summit-2026-identity-expanded-faster-than-most-programs-did/
23) Sophos. "71% of Organizations Suffered an Identity-Related Breach: The State of Identity Security 2026." Sophos Press, 2026.05.12. https://www.sophos.com/en-us/press/press-releases/2026/05/71-percent-organizations-suffered-identity-breach-state-of-identity-security-2026
24) GitGuardian. "The State of Secrets Sprawl 2026." GitGuardian Blog, 2026. https://blog.gitguardian.com/the-state-of-secrets-sprawl-2026/
25) CrowdStrike. "CrowdStrike to Acquire SGNL to Transform Identity Security for the AI Era." CrowdStrike Press, 2026.01.08. https://www.crowdstrike.com/en-us/press-releases/crowdstrike-to-acquire-sgnl-to-transform-identity-security-for-ai-era/
26) Palo Alto Networks. "Palo Alto Networks Completes Acquisition of CyberArk to Secure the AI Era." Palo Alto Networks Press, 2026.02.11. https://www.paloaltonetworks.com/company/press/2026/palo-alto-networks-completes-acquisition-of-cyberark-to-secure-the-ai-era
27) Cisco. "Cisco Announces Intent to Acquire Astrix Security." Cisco Newsroom, 2026.05.04. https://blogs.cisco.com/news/cisco-announces-intent-to-acquire-astrix-security

이 전환의 한가운데에서 주요 신원 플랫폼들이 경쟁하는데, 이번 글에서는 성격이 뚜렷하게 갈리는 다섯 곳을 살펴본다. 신원 허브의 리더 옥타(Okta), 특권 접근의 강자이자 팔로알토에 인수된 사이버아크(이디라), 엔터프라이즈의 기본 신원 인프라 마이크로소프트 엔트라 ID, 그리고 클라우드 네이티브 IAM을 대표하는 두 하이퍼스케일러 AWS IAM 아이덴티티 센터와 구글 클라우드다. 이들은 다음 하나의 질문에 대한 답을 제안하고 있다. "사람의 로그인을 관리하던 체계를, 사람이 아닌 것들이 지배하는 환경에 어떻게 다시 세울 것인가?"

## 2026년 IAM의 세 가지 큰 흐름

### 비인간 ID의 폭증 — 관리되지 않는 다수

비인간 ID는 하나의 제품군이 아니라 성격이 제각각인 여러 신원을 한데 묶어 부르는 포괄적 개념이다. 애플리케이션 간 호출에 쓰는 서비스 계정, 외부 접속용 API 키, 위임 권한을 담은 OAuth 토큰, 서버 신뢰를 증명하는 인증서, 지속적 통합·배포(CI/CD) 파이프라인의 실행 주체, 쿠버네티스 서비스 어카운트가 모두 여기에 속한다. 공통점은 대개 개발이나 초기 운영 과정에서 "일단 작동하게" 만들어졌다가, 그 뒤로 아무도 회수하지 않은 채 영구 권한과 함께 남는다는 데 있다.

문제는 이 신원들이 수적으로 방대할 뿐 아니라, 그 대부분이 아무 관리 없이 방치된다는 데 있다. 사이버아크 조사에 따르면 특권 접근을 완전히 적시(JIT, Just-in-Time) 방식으로 전환한 조직은 전 세계 기업의 1%에 불과하고, 91%는 특권 접근의 절반 이상을 상시 켜진(always-on) 상태로 유지한다.[28] 필요할 때만 잠깐 열렸다 닫혀야 할 권한이, 쓰지 않는 순간에도 계속 열려 있는 것이다. 여기에 상당수의 비인간 ID가 문서화조차 되지 않은 채 오래 방치된다는 사실이 더해지면, 한 번 유출된 자격 증명이 오래도록 유효하게 남게 되는데, 이 문제의 무게는 시장 규모로도 드러난다. 비인간 ID 접근 관리 시장은 2024년 약 94억 5,000만 달러에서 2030년 약 187억 달러로, 연평균 11.9% 성장할 것으로 전망된다.[29] 방화벽을 아무리 높여도, 이미 방화벽 내부에서 과도한 권한을 쥔 채 잊힌 신원이 수만 개라면 경계선은 의미를 잃는다.

---

28) CyberArk. "New Study: Only 1% of Organizations Have Fully Adopted Just-in-Time Privileged Access as AI-Driven Identities Rapidly Increase." CyberArk Press, 2026. https://www.cyberark.com/press/new-study-only-1-of-organizations-have-fully-adopted-just-in-time-privileged-access-as-ai-driven-identities-rapidly-increase/

29) MarketsandMarkets. "Non-Human Identity (NHI) Access Management Market." MarketsandMarkets Press Release, 2026.06. https://www.marketsandmarkets.com/PressReleases/non-human-identity-nhi-access-management.asp

## AI 에이전트 — 완전히 새로운 유형의 신원

기존의 비인간 ID들에 비해, AI 에이전트는 성격이 다르다. 서비스 계정은 정해진 일을 반복하지만, 에이전트는 상황에 따라 스스로 판단해 도구를 고르고 데이터에 접근하며 다음 행동을 결정하는 등 한자리에 고정돼 있지 않다. 인간처럼 능동적으로 움직이되, 인간처럼 로그인하지는 않는다. 6월호에서 다룬 모델 컨텍스트 프로토콜(MCP)의 확산은 이 문제를 키웠는데, 깃가디언이 MCP 관련 설정 파일에서만 2,117개의 유효한 자격 증명을 발견했다는 사실은, 에이전트가 쥔 열쇠가 얼마나 허술하게 흩어져 있는지를 보여준다.

그럼에도 대다수 조직은 아직 무엇이 있는지 파악하는 단계에 머물러 있다. 시스코가 아스트릭스 인수를 발표하며 인용한 자사 AI 준비도 지수에서, AI 에이전트의 행동을 안전하게 통제할 수단을 갖춘 조직은 24%에 그쳤고, 옥타가 인용한 조사에서는 69%의 조직이 직원의 미승인 생성형 AI 도구 사용을 의심하거나 확인했다고 답했다.[30] 사고의 조짐은 이미 나타나고 있는데, 정작 그 에이전트가 누구이고 무엇을 할 수 있는지 파악한 조직은 소수에 불과하다.

가트너 서밋에서는 신원 관리의 대상이 사람을 넘어 워크로드와 에이전트, 자격 증명으로 급격히 확장되고 있다는 진단이 이어졌다. 신원의 관측 대상 자체가 사람에서 에이전트로 옮겨 가는 것이고, 벤더들의 대응은 한 방향으로 수렴한다. 에이전트를 사람 계정에 얹거나 익명의 서비스 계정으로 두지 말고, 소유자와 수명과 권한 범위를 가진 독립된 신원으로 등록하라는 것이다.

## 신원으로 옮겨 간 보안의 축

보안 관련 벤더들의 2026년 상반기 인수들이 공통으로 겨냥한 것은 단순한 규모 확장이 아니라, 신원을 보안 플랫폼의 한가운데로 끌어들이는 통합이다. 연속 인가, 특권 접근 관리, 비인간 ID 발견 같은 신원 역량이 네트워크·엔드포인트 보안과 한 플랫폼으로 합쳐지고 있다.

이는 3월호에서 다룬 보안 벤더들의 플랫폼화 전략과 곧바로 이어진다. 3월호에서는 "위협을 어떻게 탐지하고 대응하는가"를 사고 이후의 시각에서 다뤘다면, 이번 호에서는 "누가 접근하는가"를 다루며 사고 이전의 예방에 중점을 두고 있으며, 두 영역은 이제 벤더 차원에서 하나로 합쳐지고 있다.

30) Okta. "Okta Secures the Agentic Enterprise with New Tools for Discovering and Mitigating Shadow AI Risks." Okta Press, 2026.02.12. https://www.okta.com/newsroom/press-releases/okta-secures-the-agentic-enterprise-with-new-tools-for-discovering-and-mitigating-shadow-ai-risks/

보안팀과 IAM팀이 서로 다른 부서에서 각자의 도구로 움직이는 구조에서는 이 흐름이 어긋나기에, 신원 기반 위협 탐지·대응(ITDR)이 보안 운영의 일부로 편입되도록 필요한 경우 조직을 재편해야 한다.

## 벤더 심층 분석: 신원 통제 역량의 비교

신원과 관련한 다섯 벤더는 아래 표 1과 같이 다섯 가지의 축으로 비교한다.

1) 발견 — 관리되지 않는 비인간 ID와 섀도 에이전트를 가시화하는가
2) 에이전트 신원 등록 — 에이전트를 소유자와 수명을 가진 독립 ID로 관리하는가
3) 적시·최소 권한 — 상시 특권을 제거하고 필요할 때만 부여하는가
4) 동적 인가 — 접근 시점의 상황에 따라 권한을 실시간으로 판정하는가
5) 다중 환경 적용성 — 특정 클라우드나 생태계에 종속되지 않는가?

앞의 셋이 저마다 에이전트 전용 신원 제품을 내놓고, 두 하이퍼스케일러(AWS·GCP)는 워크로드마다 짧은 수명의 암호학적 신원을 발급하는 개방 표준인 스파이피(SPIFFE, Secure Production Identity Framework For Everyone)를 기반으로 삼는다. 다만 AWS가 기존 IAM 강화에 무게를 둔다면, GCP는 그 표준 위에 에이전트 전용 신원을 퍼스트클래스로 올린 점이 다르다.

| 구분 | 옥타 | 사이버아크/이디라 | 엔트라 ID | AWS | GCP |
|---|---|---|---|---|---|
| 핵심 포지셔닝 | 신원 허브 리더 | 특권 접근(PAM) 리더 | 엔터프라이즈 기본 신원 | 클라우드 네이티브 IAM | 에이전트 신원 특화 |
| 발견 | 에이전트 발견(ISPM) | AI 기반 접근 경로 발견 | 워크로드 신원 인벤토리 | IAM Access Analyzer | Agent Gateway 중앙 경유 |
| 에이전트 신원 등록 | Okta for AI Agents | 특권 사용자로 취급 | 엔트라 에이전트 ID | 워크로드 신원 페더레이션 | Agent Identity(SPIFFE) |
| 적시·최소 권한 | 부분 지원 | 제로 스탠딩 특권 + JIT | 조건부 접근 연계 | 단기 자격 증명 | 24시간 인증서 + PAB |
| 동적 인가 | 정책 + SIEM 스트리밍 | AI 기반 정책 자동화 | 조건부 접근(강점) | IAM 정책 + 개방 표준 | IAP + Unified Access Policy |
| 다중 환경 적용성 | 높음(Any IdP) | 높음 | 애저 중심 | AWS 중심 | GCP 중심(SPIFFE 표준) |
| 약점 | 신제품 초기·신뢰 회복 | 인수 통합·워크포스 IAM | MS 종속·라이선스 복잡 | 에이전트 전용 부재 | IAP·UAP 프리뷰/예정 |

**표 1** 2026년 IAM/ID 5개 벤더의 신원 통제 역량 비교

### 옥타 — 모든 에이전트에는 신원이 필요하다

옥타는 8,200개가 넘는 사전 통합 카탈로그를 앞세운 신원 허브의 리더다. SaaS 생태계 어디에 붙든 옥타를 거쳐 로그인하도록 설계된 통합의 폭, 그것이 옥타의 해자(moat)였고, 2026년 옥타는 이 자리를 에이전트 시대로 확장하겠다고 선언했다. 2026년 4월 정식 출시된 'Okta for AI

Agents'에서는 에이전트를 유니버설 디렉터리에 비인간 ID로 등록해 소유자를 지정하고 수명을 관리하며, 에이전트의 도구 호출과 인증 결정, 접근 시도를 보안 정보·이벤트 관리(SIEM) 시스템으로 실시간으로 흘려보낸다.[31]

한편, 2026년 2월 공개한 에이전트 발견(Agent Discovery) 기능은 신원 보안 태세 관리(ISPM)의 일부로 작동하는데, 미승인 에이전트를 감지하고, OAuth 동의 흐름을 추적하며, 발견된 섀도 에이전트를 관리 대상 신원으로 전환해 등록한다. 옥타가 인용한 조사에서 69%의 조직이 직원의 미승인 생성형 AI 사용을 의심하거나 확인했다는 수치가 있는데, 이는 이 기능이 필요한 사각지대의 크기를 보여준다.

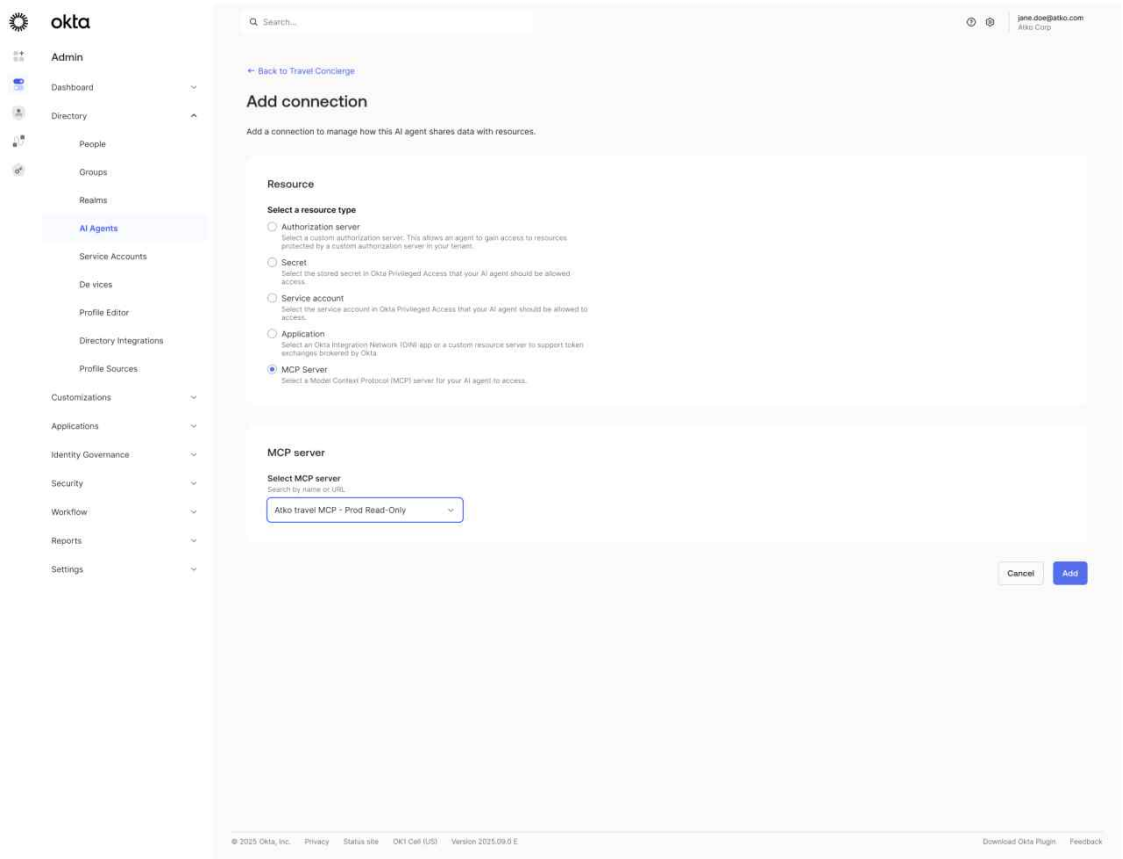


**그림 6** Okta for AI Agents — 에이전트를 비인간 ID로 등록하고 섀도 AI를 발견하는 화면 (출처: Okta)

31) Okta. "Okta Showcase 2026: Securing Every Identity, Including AI Agents." Okta Press, 2026.04.30. https://www.okta.com/newsroom/press-releases/showcase-2026/

반면, Okta for AI Agents가 아직 신제품이어서 대규모 프로덕션 레퍼런스가 적은데, 비인간 ID 관리가 전용 제품이 아니라 기존 신원 보안 플랫폼의 확장으로 제공되는 만큼, 전문 도구에 비해 깊이가 부족할 수 있다. 그리고 엔터프라이즈 플랜 가격도 높은 편이어서 중소 조직에는 진입 장벽이 되고 있으며, 2023년 지원 시스템 침해 이후의 신뢰 회복도 여전히 과제로 남아 있다.

## 사이버아크(이디라) — 특권 접근을 AI 에이전트로 확장하다

사이버아크는 특권 접근 관리(PAM)의 오랜 리더로, 시스템을 통째로 좌우할 수 있는 특권 계정을 격리하고 감시하며 자격 증명을 주기적으로 교체하는 것을 핵심 역량으로 삼아 왔다. 2026년 2월 팔로알토 네트웍스가 인수를 완료하면서 사이버아크의 특권 접근 역량은 네트워크와 엔드포인트를 아우르는 팔로알토 플랫폼의 신원 축으로 자리 잡았고, 그 결실이 5월 12일 출시된 차세대 플랫폼 이디라다.

이디라는 AI 기반으로 신원과 권한, 접근 경로를 지속적으로 발견하고, 제로 스탠딩 특권(Zero Standing Privileges, 평상시에는 아무 권한도 없다가 필요할 때만 부여하는 원칙)과 적시 접근을 강제하며, AI 정책으로 컴플라이언스 준수를 수작업 점검에서 자동화된 거버넌스로 바꾼다. 이는 머신과 AI 신원이 인간을 압도하고 상시 특권이 방치되는 현실을 정면으로 겨냥한 설계로, AI 에이전트가 상시 특권을 쥔 채 탈취될 경우 사람과 비교할 수 없이 커지는 피해를 막기 위한 것이다.

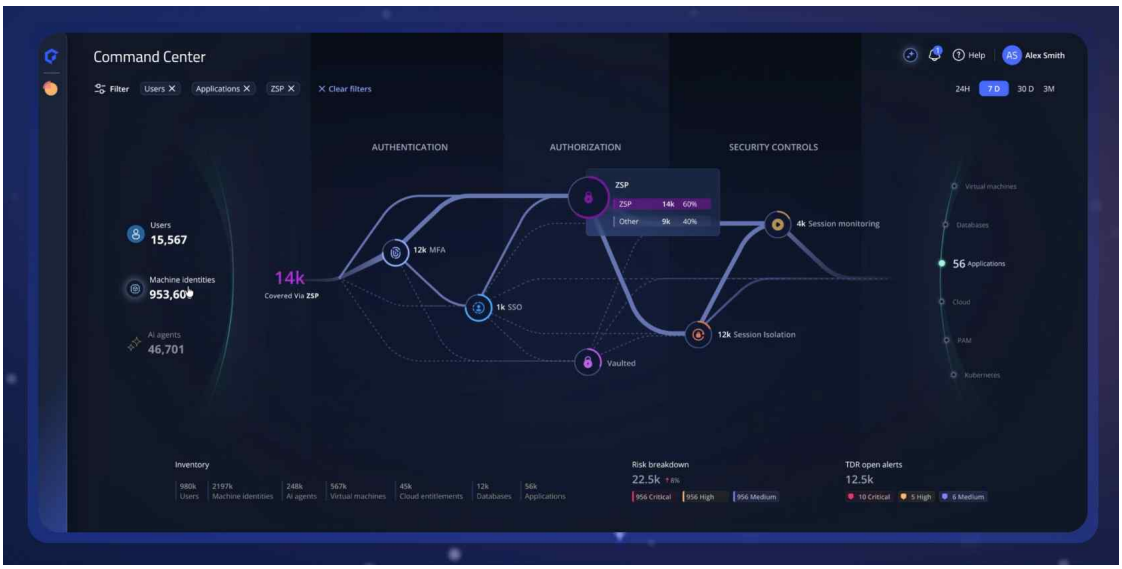


**그림 7** 사이버아크 이디라 커맨드 센터 — 머신·에이전트 신원과 ZSP(제로 스탠딩 특권) 기반 인증·인가·통제 흐름 (출처: 팔로 알토 네트웍스)

다만 인수 직후라 기존 고객이 이디라로 넘어가는 마이그레이션 경로와 팔로알토 플랫폼 기준의 가격 재편이 아직 불확실하다. 특권 접근(PAM)에는 강하지만 SSO·MFA 중심의 워크포스 IAM에서는 옥타·엔트라에 밀리고, PAM 특유의 구현 복잡성과 비용도 오랜 과제다.

## 마이크로소프트 엔트라 ID — 조건부 접근을 에이전트로 넓히다

엔트라 ID는 마이크로소프트 365와 애저를 쓰는 조직에 사실상 기본으로 깔리는 엔터프라이즈 신원 인프라로서, 별도 도입을 고민할 필요 없이 이미 그 자리에 있는 경우가 많다. 엔트라의 무기는 조건부 접근이라는 정책 엔진으로, 사용자·기기·위치·위험도를 조합해 접근을 실시간으로 판정하는 제로 트러스트의 핵심 장치다. 2026년, 마이크로소프트는 이 조건부 접근을 에이전트로 확장했다.

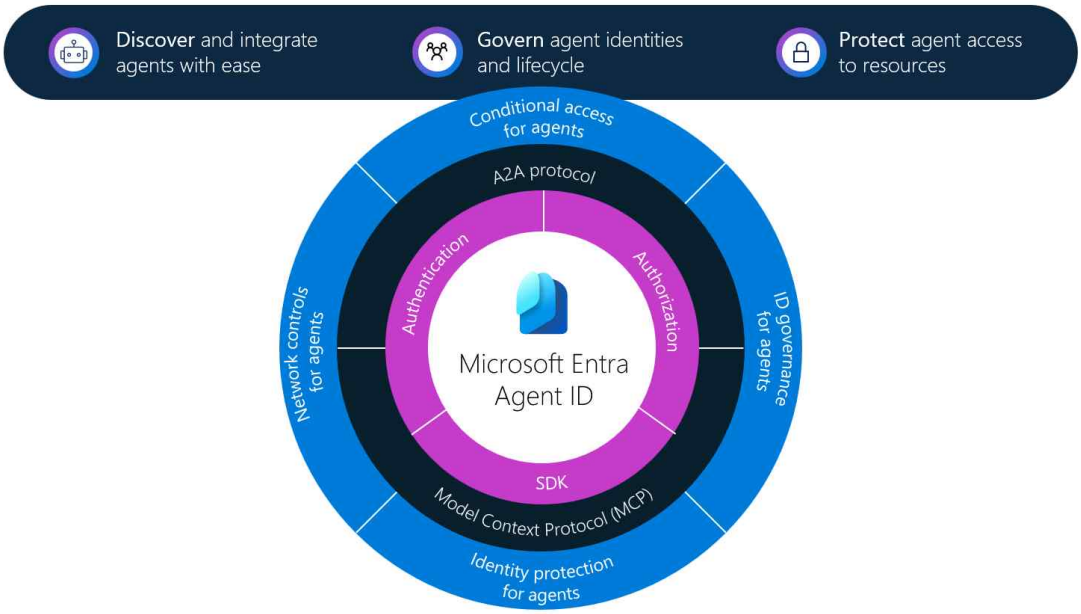


**그림 8** 마이크로소프트 엔트라 에이전트 ID — 에이전트 신원의 발견·거버넌스·보호와 조건부 접근 체계 (출처: 마이크로소프트)

엔트라 에이전트 ID는 2026년 상반기에 정식 출시되었는데. AI 에이전트를 퍼스트클래스 신원으로 등록하고, 인간·앱·기기에 적용하던 것과 동일한 조건부 접근 정책을 에이전트에도 적용한다. 적응형 접근 정책, 실시간 위험 감지, 라이프사이클 관리가 하나의 제로 트러스트 틀 안에서 에이전트에까지 적용된다는 것이 핵심이고[32], AWS 베드락 같은 서드파티 에이전트도 엔트라 인증 소프트웨어 개발

키트(SDK)나 워크로드 신원 페더레이션으로 온보딩할 수 있다. 이와 같은 조건부 접근 최적화 에이전트는 정책 작업을 43% 더 빠르고 48% 더 정확하게 처리한다고 마이크로소프트는 밝혔다.[33]

반면, 마이크로소프트 생태계에 종속돼 비애저·비M365 환경에서는 유용성이 제한되고, 코파일럿 생태계 바깥의 에이전트에 대한 가시성에는 한계가 있다. 에이전트 ID가 아직 신규 기능이어서 프로덕션 사례가 더 쌓여야 하며, 라이선스 체계가 복잡해 비용을 예측하기 어렵다는 점도 자주 지적된다.

## AWS IAM 아이덴티티 센터 — 신제품 대신 기존 IAM을 강화하다

AWS IAM 아이덴티티 센터는 AWS 워크로드를 중심으로 다수 계정에 대한 접근을 통합 관리하는 클라우드 네이티브 IAM이다. 2026년 2월 멀티 리전 복제를 도입해, 기본 리전에 장애가 나도 다른 리전에서 접근을 복원할 수 있게 했고[34], IAM 액세스 어낼라이저는 조직 전체의 정책을 평가해 쓰이지 않는 권한을 표시하고, 클라우드트레일(CloudTrail) 기록을 근거로 최소 권한을 추천한다.

AWS의 접근법은 다른 벤더와 방향이 다른데, 에이전트 전용 신원 제품을 서둘러 내놓기보다 기존 IAM을 강화하고 표준을 다지는 쪽을 택했다. SPIFFE와 IETF(Internet Engineering Task Force)의 윔지(WIMSE, Workload Identity in Multi-System Environments)를 채택해, 멀티클라우드 환경에서 정적 API 키를 단기 자격 증명으로 대체하고 있다. 베드락 에이전트코어와 연계하면 에이전트의 생명주기 안에서 IAM 정책이 자동 적용되는 등, 새 제품을 내세우는 대신 표준과 단기 자격 증명으로 비인간 ID 문제를 보수적이지만 견고하게 해결한다.

다만 그 신중함에 따른 대가로, 옥타나 엔트라와 달리 에이전트 전용 신원 거버넌스 제품이 없어, 같은 수준의 통제를 원한다면 직접 구축해야 한다. AWS 생태계에 최적화된 만큼 멀티클라우드 신원 관리에서는 두 벤더에 밀리고, 비인간 ID 관리의 깊이도 전문 벤더에는 못 미친다는 평이다.

---

32) Microsoft. "What is Microsoft Entra Agent ID." Microsoft Learn, 2026. https://learn.microsoft.com/en-us/entra/agent-id/what-is-microsoft-entra-agent-id
33) Microsoft. "Four Priorities for AI-Powered Identity and Network Access Security in 2026." Microsoft Security Blog, 2026.01.20. https://www.microsoft.com/en-us/security/blog/2026/01/20/four-priorities-for-ai-powered-identity-and-network-access-security-in-2026/
34) AWS. "AWS IAM Identity Center Multi-Region AWS Account Access and Application Deployment." AWS What's New, 2026.02. https://aws.amazon.com/about-aws/whats-new/2026/02/aws-iam-identity-center-multi-region-aws-account-access-and-application-deployment/

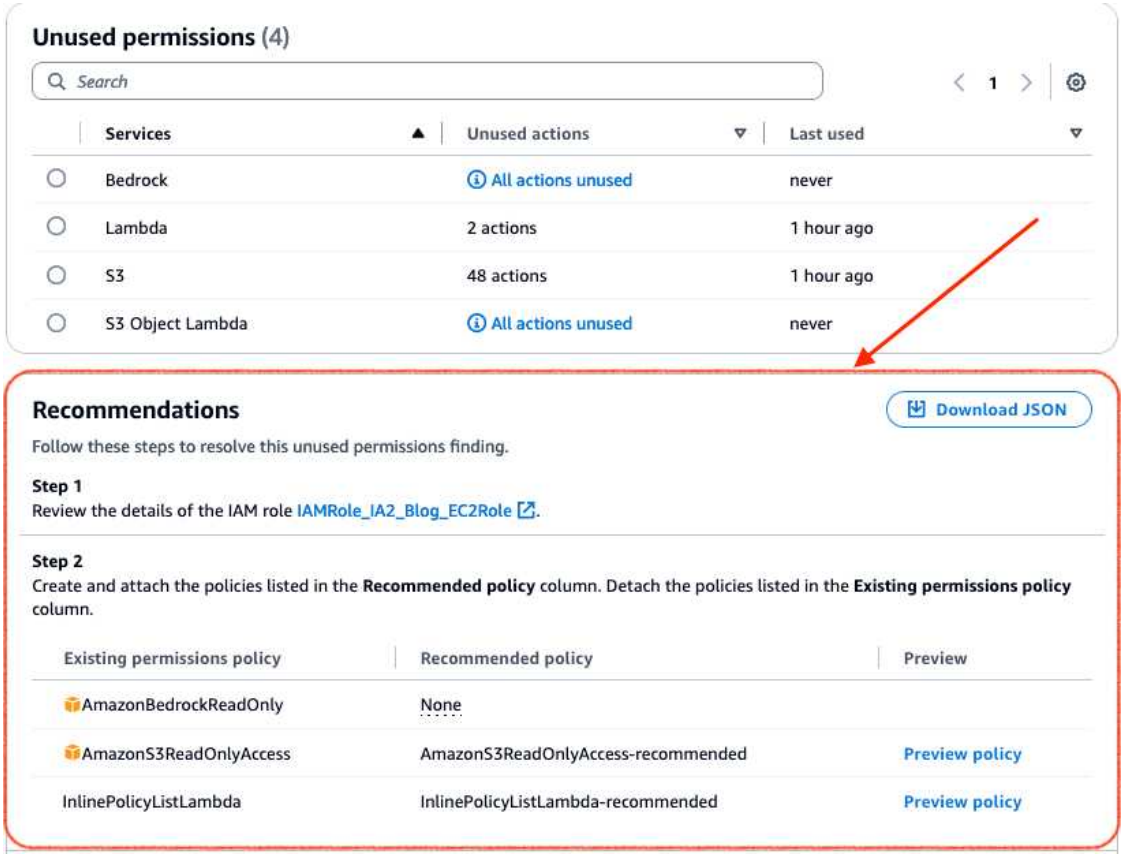


그림 9 AWS IAM 액세스 어낼라이저 — 미사용 권한 표시와 최소 권한 추천 (출처: AWS)

## 구글 클라우드 — 표준 위에 세운 에이전트 전용 신원

구글 클라우드는 2026년 5월 7일 구글 클라우드 넥스트에서 에이전트 신원을 전면에 내세웠다.[35] 핵심은 에이전트 아이덴티티로, AI 에이전트에 사람이나 서비스 계정과 구분되는 퍼스트클래스 신원을 부여한다. SPIFFE를 기반으로 에이전트마다 고유한 암호학적 ID를 발급하고, X.509 인증서를 자동으로 프로비저닝해 24시간마다 갱신하는데, 발급된 액세스 토큰은 에이전트의 X.509 인증서에 암호학적으로 묶여, 탈취되더라도 다른 곳에서 쓰이지 못한다. 공유 서비스 계정과 달리 각 에이전트 신원은 위장이 불가능하다는 점이 설계의 핵심이다.[36]

35) Google Cloud. "What's New in IAM: Security, Governance, and Runtime Defense." Google Cloud Blog, 2026.05.07. https://cloud.google.com/blog/products/identity-security/whats-new-in-iam-security-governance-and-runtime-defense

에이전트에 신원을 부여한 다음은 그 신원이 무엇에 닿을 수 있는지를 정하는 통제의 단계로, 구글 클라우드는 이를 게이트웨이와 정책으로 풀어낸다. 에이전트 게이트웨이는 정식 출시되어, 모든 에이전트 간·에이전트-도구 연결을 중앙 경로로 강제하고 미승인 엔드포인트 접근을 차단한다. 여기에 Identity-Aware 프록시가 결합해(프리뷰) 에이전트 신원과 MCP 컨텍스트를 근거로 세밀한 접근을 판정하며, 곧 나올 통합 접근 정책(UAP)은 민감한 작업에 사람의 승인(human-in-the-loop)을 요구하는 규칙까지 담는다.

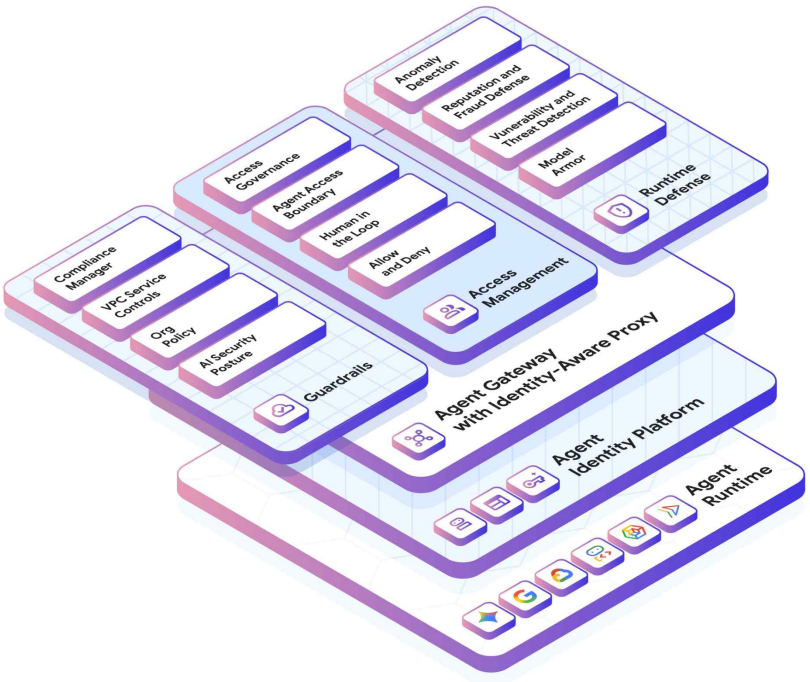


그림 10 구글 클라우드의 AI 에이전트 보안·거버넌스 계층 구조 (출처: 구글 클라우드)

앞의 AWS가 기존 IAM 강화에 무게를 뒀다면, 구글 클라우드는 같은 SPIFFE 표준 위에 에이전트 전용 신원을 정면으로 세운 점이 다르다. 다만 에이전트 아이덴티티와 에이전트 게이트웨이는 정식 출시됐으나 Identity-Aware 프록시는 프리뷰, UAP는 출시 예정 단계여서 통제 스택이 아직 완결되지 않았으며, 클라우드 아이덴티티의 엔터프라이즈 신원 저변도 옥타·엔트라에 비해 얕고, GCP 생태계 밖에서 단독으로 쓰기에는 통합 부담이 남아 있다.

36) Google Cloud. "Agent Identity Overview." Google Cloud IAM Documentation, 2026. https://docs.cloud.google.com/iam/docs/agent-identity-overview

## 맺으며: CIO·CTO를 위한 신원 전략 제언

2026년의 신원 전략은 새 인증 수단의 도입이 아니라, 보이지 않던 다수를 관리 대상으로 끌어들이는 데서 시작한다. 사람의 로그인만 관리하던 체계로는, 사람보다 수십 배 많은 비인간 ID와 스스로 판단하는 에이전트를 감당할 수 없지만, 그렇다고 모든 조직이 지금 당장 신원 플랫폼을 새로 들여야 하는 것은 아니다. AI 에이전트가 이미 사내 데이터에 접근하고 있거나, 서비스 계정·API 키가 통제 없이 늘고 있거나, 규제가 신원 체계의 재정비를 요구한다면 서둘러야겠지만, 그렇지 않다면 발견과 인벤토리부터 시작해도 늦지 않다. 신원 체계를 다시 세워야 하는 조직에 다음 네 가지를 권한다.

비인간 ID의 인벤토리부터 수행하라. 보이지 않는 것은 관리할 수 없기에, 대부분의 조직은 자사 환경에 비인간 ID가 몇 개인지조차 알지 못한다. 서비스 계정, API 키, OAuth 토큰, 인증서를 전수 조사하고 각각의 권한 수준과 마지막 사용 시점, 소유자를 파악하는 것이 첫 단계이다. 옥타의 신원 보안 태세 관리 기능이나 전문 비인간 ID 발견 도구가 이 작업을 자동화한다.

AI 에이전트를 독립된 신원으로 등록하고 라이프사이클을 관리하라. 에이전트가 조직의 도구와 데이터에 접근한다면 고유한 신원이 있어야 하고, 사람 계정을 공유하거나 익명의 서비스 계정으로 두어서는 안 된다. Okta for AI Agents, 엔트라 에이전트 ID, 구글 클라우드 에이전트 아이덴티티 가운데 기존 인프라에 맞는 경로를 고르되, "누가 이 에이전트를 소유하는가", "수명은 언제까지인가", "접근 범위는 어디까지인가" 세 질문에 명시적으로 답할 수 있어야 한다.

적시·최소 권한을 사람과 기계 모두에 적용하라. 에이전트가 탈취되면 사람보다 빠르게 권한을 악용하기 때문에 상시 권한은 에이전트 시대에 훨씬 위험하다. 영구 자격 증명을 단기 토큰으로 바꾸고 작업이 끝나면 즉시 권한을 회수하는 적시 모델을, 인간 신원과 비인간 ID 모두에 확대하는 것을 권한다.

보안팀과 IAM팀의 사일로를 해체하라. 조직 내부에서 두 팀이 별도로 움직이는 구조는 보안과 신원이 하나로 합쳐지고 있는 흐름과 어긋난다. 신원 기반 위협 탐지·대응을 보안 운영 센터(SOC)의 워크플로에 통합하고, 비인간 ID 관리를 보안 운영의 일부로 편입하기를 권한다. 여기에 국내 조직에는 고려할 시점이 하나 더 있는데, 공공 클라우드 보안 인증(CSAP)이 국정원 단일 체계로 개편되면서(2027년 7월 시행), 정보보호 관리 체계(ISMS-P) 등 유사 인증과 중복되는 항목은 검증에서 면제되는 방향으로 정비된다.[37] 공공 클라우드 공급사뿐 아니라 공공기관에 납품하는 솔루션 기업들도

37) ZDNet Korea. "CSAP, 국정원 단일 체계로 개편." ZDNet Korea, 2026.04.20. https://zdnet.co.kr/view/?no=20260420130424

신원 아키텍처를 규정에 맞게 재설계해야 한다.

이와는 반대로 피해야 할 것도 분명한데, 인벤토리 없이 도구부터 들이는 일, 에이전트에 사람 계정을 물려 익명으로 굴리는 일, "에이전트 ID 지원" 발표만 보고 신원 관리가 저절로 된다고 기대하는 일이다. 셋 다 관리 대상을 밝히는 대신 사각지대를 넓혀 역효과가 일어난다.

공격자는 더 이상 뚫고 들어오지 않는다. 그들은 로그인한다. 2026년 신원 전략의 출발점은 누가 로그인하는지를, 그것이 사람이든 에이전트든, 조직이 먼저 아는 것이다.

**부록**

| 구분 | 옥타 | 사이버아크/이디라 | 엔트라 ID | AWS | GCP |
|---|---|---|---|---|---|
| **핵심 포지셔닝** | 신원 허브 리더 | 특권 접근(PAM) 리더 | 엔터프라이즈 기본 신원 | 클라우드 네이티브 IAM | 에이전트 신원 특화 |
| **발견** | 에이전트 발견(ISPM) | AI 기반 접근 경로 발견 | 워크로드 신원 인벤토리 | IAM Access Analyzer | Agent Gateway 중앙 경유 |
| **에이전트 신원 등록** | Okta for AI Agents | 특권 사용자로 취급 | 엔트라 에이전트 ID | 워크로드 신원 페더레이션 | Agent Identity(SPIFFE) |
| **적시·최소 권한** | 부분 지원 | 제로 스탠딩 특권 + JIT | 조건부 접근 연계 | 단기 자격 증명 | 24시간 인증서 + PAB |
| **동적 인가** | 정책 + SIEM 스트리밍 | AI 기반 정책 자동화 | 조건부 접근(강점) | IAM 정책 + 개방 표준 | IAP + Unified Access Policy |
| **다중 환경 적용성** | 높음(Any IdP) | 높음 | 애저 중심 | AWS 중심 | GCP 중심(SPIFFE 표준) |
| **약점** | 신제품 초기·신뢰 회복 | 인수 통합·워크포스 IAM | MS 종속·라이선스 복잡 | 에이전트 전용 부재 | IAP·UAP 프리뷰/예정 |

**표 2** 2026년 IAM/ID 5개 벤더의 신원 통제 역량 비교