AI Live + Labs

Agentic AI 보안 거버넌스 글로벌 동향

Architecting a secure agentic ecosystem with Gemini Enterprise Agent Platform

Google Cloud

Agenda

01 **Agent** 거버넌스 개괄

02 **Gemini Enterprise Agent Platform** - 보안 및 거버넌스

03 구글클라우드의 **AI & Agents** 보안 강화 방식

기업내 에이전트 거버넌스

No one really knows what an AI agent is

https://techcrunch.com/2025/05/12/even-a16z-vcs-say-no-one-really-knows-what-an-ai-agent-is/

Agent, Assistant & Bot

에이전트란 자신의 환경을 인지하고
그 환경에 작용할 수 있는 모든 것을
의미합니다

https://huyenchip.com//2025/01/07/agents.html

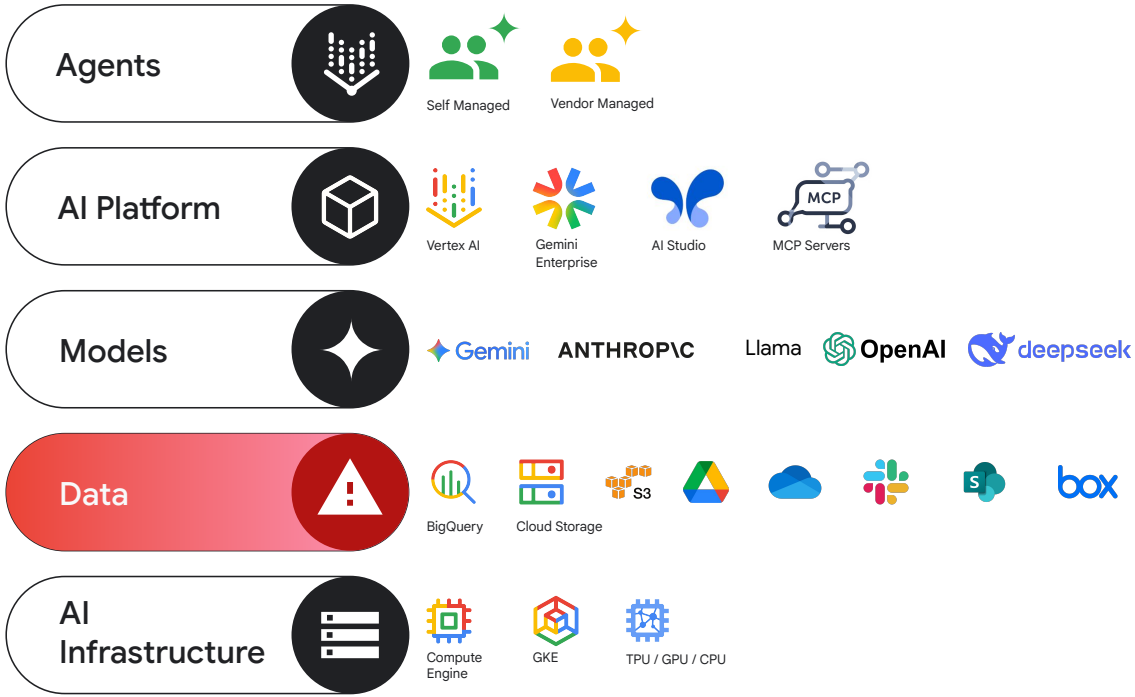
에이전트는 **LLM**이 자신의 프로세스와 도구 사용을 동적으로 지휘하며, 작업을 완수하는 방식에 대한 제어권을 유지하는 시스템입니다

https://www.anthropic.com/engineering/building-effective-agents

AI 에이전트는 사용자를 대신하여 목표를 추구하고 작업을 완수하기 위해 AI를 활용하는 소프트웨어 시스템입니다.

https://cloud.google.com/discover/what-are-ai-agents

Agents interact with the entire AI stack



Agents can amplify risks across the AI stack

Extracting meaning from text can be hard.

Emphasis changes everything.

I *never* said she stole my money.
Someone else may have said it, but it wasn't me.

I **never** said she stole my money.
I didn't make the claim at any point in time.

I never **said** she stole my money.
I may have implied or thought it, but didn't say it.

I never said **she** stole my money.
Someone else may have stolen it, but it wasn't her.

I never said she **stole** my money.
She may have borrowed or been given it.

I never said she stole **my** money.
She stole someone else's money.

I never said she stole my **money.**
She may have stolen something else.

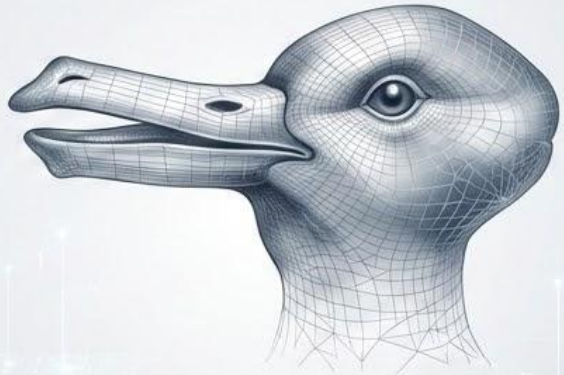
If text with similar ambiguity is in inputs (e.g. prompts), how can we be sure the agent is interpreting as intended?

Text is messy, language models are messy

https://x.com/bsacash/status/1793624024226168955

Context is the new perimeter.

Systems need to consider context



Avoid All Ducks...
This may not be a duck !!!

Agents everywhere, expect bumps, Embrace the journey

에이전트의 보편화
사용되는 모든 도구 내 사전 패키징된 AI 내재화

단일 봇에서 팀 단위 협업 체계로 전환
개별 봇 운영에서 유기적으로 조율되는 멀티 에이전트 팀으로 전환

거버넌스 및 통제 역량 성숙
안전성 확보를 위한 거버넌스·감사·관리 도구의 고도화

새로운 업무 워크플로우 정착
에이전트 기반 작업 방식이 표준 업무 수행 방식으로 자리매김

도입기 시행착오의 필연성
초기 난관 및 과잉 설계는 학습과 적응을 위한 필수적 과정임

3대 핵심 시사점

에이전트에 대한 과도한 기대나 조기 배제 경계

에이전트는 하나의 수단에 불과하며, 적용 영역에 따라 최적의 효용을 발휘하거나 불필요할 수 있음.

에이전트 대응 플랫폼 인프라 구축 및 준비

에이전트 빌더 및 운영자를 위한 플랫폼 역량 평가 필요 (평가·레지스트리 서비스 필수 점검).

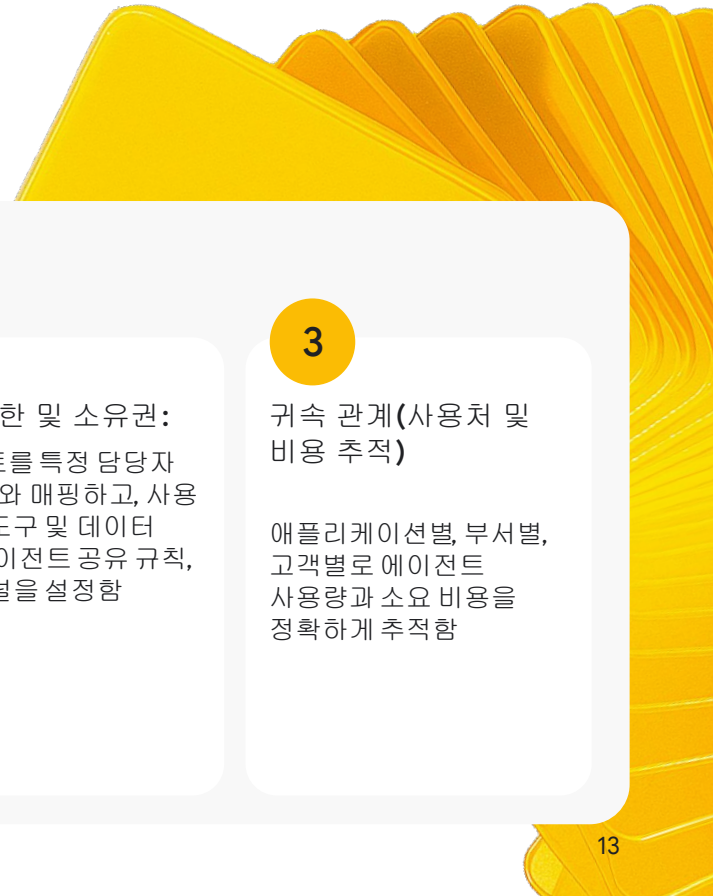
아키텍처 패턴의 고도화 및 진화

에이전틱(Agentic) 아키텍처 도입에 따라 새로운 시스템 패턴 및 사용자 상호작용 방식 대두

가시성 (Visibility)

도메인 내 모든 에이전트를 모니터링할 수 있는 단일 소스(단일 진실 공급원)가 필요

에이전트가 어떤 기술로 개발되었든, 혹은 어떤 서비스에 적용되어 작동하든 상관없이 관리자가 도메인 전체 에이전트 현황을 실시간으로 완벽하게 파악할 수 있어야 합니다.



Key requirements:

1

통합 탐색(범용적 발견):

메인 내에서 작동하는 모든 에이전트의 실시간 카탈로그 제공

2

접근 권한 및 소유권:

에이전트를 특정 담당자(소유주)와 매핑하고, 사용 가능한 도구 및 데이터 범위, 에이전트 공유 규칙, 노출 채널을 설정함

3

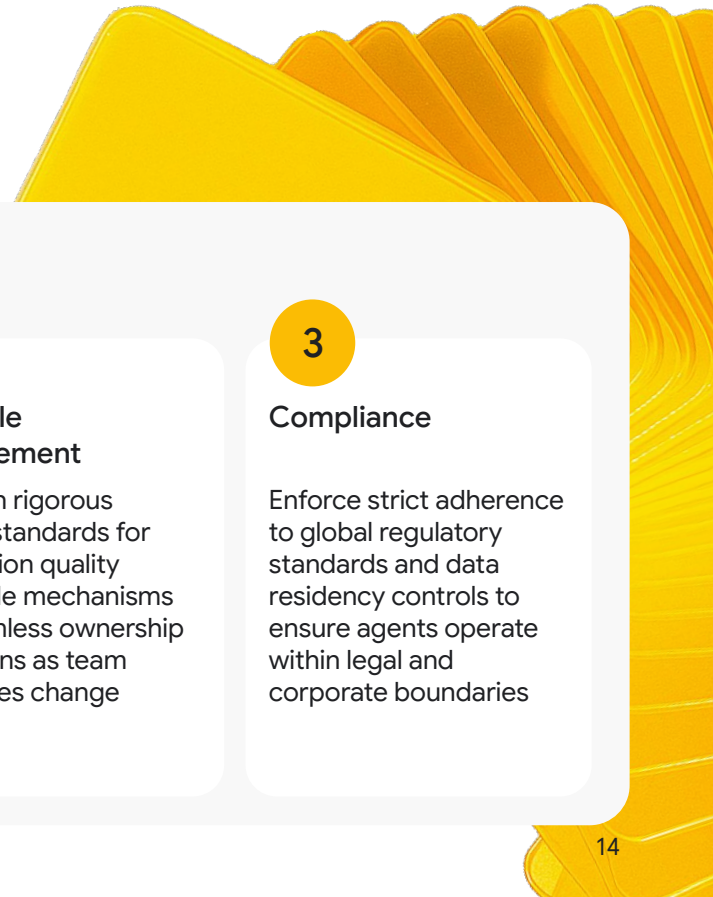
귀속 관계(사용처 및 비용 추적)

애플리케이션별, 부서별, 고객별로 에이전트 사용량과 소요 비용을 정확하게 추적함

통제 (Control)

확률적 시스템을 둘러싼 결정론적 가드레일

관리자는 에이전트의 권한을 실제 사용자의 접근 권한과 일치시키고, 프로덕션 수명주기 및 소유권 이전을 관리하며, 규제 및 데이터 주권 (레지던시) 표준 준수를 강제할 수 있는 역량이 필요합니다.



Key requirements:

1

Identity

Ensure that an agent's data access and tool usage are strictly aligned with the permissions and information boundaries of the human user it is assisting

2

Lifecycle Management

Establish rigorous testing standards for production quality alongside mechanisms for seamless ownership transitions as team structures change

3

Compliance

Enforce strict adherence to global regulatory standards and data residency controls to ensure agents operate within legal and corporate boundaries

보안 (Security)

선제적 위험 탐지 및 개입을 통한 에이전트 보안 확보

관리자는 에이전트의 통신 방식에 대한 보안 경계를 정의하고, 위협이 가시화되는 순간 이를 감지하며, 필요시 즉각 에이전트 실행을 중단할 수 있는 역량이 필요합니다.

Key requirements:

1

네트워크 및 정책 **(Network and policy)**

에이전트가 감사 가능한 안전한 채널을 통해서만 인가된 시스템과 통신하도록 보장하는 보안 경계 및 상호작용 규칙을 수립합니다.

2

위협 탐지 **(Threat detection)**

에이전트의 페이로드(입출력 데이터)와 행동 패턴을 지속적으로 모니터링하여 프롬프트 인젝션(Prompt Injection), 민감 데이터 유출, 악의적 의도를 실시간으로 식별합니다.

3

비정상/탈주 에이전트 즉각 차단 **(Stop a rogue Agent)**

보안 임계값이 위반되거나 비정상(Rogue) 상태가 식별되는 즉시 에이전트 실행을 강제 종료하는 서킷 브레이커(Circuit Breaker, 비상 정지 메커니즘)를 구축합니다.

에이전트 거버넌스 (Agent Governance)

Open foundations. Integrated stack. Enterprise trust.

기존 인프라 활용 (Leverage existing infra)

이미 신뢰하고 검증된 기존 인프라를 확장하여, 최신 애플리케이션과 함께 AI 에이전트를 안정적으로 호스팅합니다.

에이전트 대우 (Agents as first-class)

현재 적용 중인 엄격한 보안, 신원 관리 (Identity), 컴플라이언스 기준을 에이전트에도 동일하게 적용하여 차세대 지능형 시스템으로의 원활한 진화를 보장합니다.

Open platform

확장 가능한 오픈소스 기반 기술을 활용해 AI 에이전트를 관리하며, SAIF(Secure AI Framework), 표준 프로토콜, Envoy 등 개방형 에코시스템과 협력하여 보안을 강화합니다.

Single pane of glass

마이크로서비스부터 자율 에이전트 군집 (Autonomous Fleet)에 이르기까지 모든 인프라와 서비스를 한곳에서 관제하고 통제합니다.

Integrated stack

Google의 수직 통합형 AI 스택은 내부적 대규모 도입을 이끌며, 에이전트 기능의 고도화를 가속화합니다.

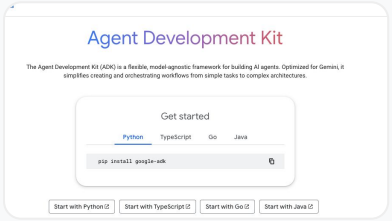
Scalable growth

에이전트를 도구 및 데이터와 연결하는 모범 경로(Golden Paths)를 제공하여, 개발자가 기존 엔터프라이즈 역량을 손쉽게 탐색하고 재사용할 수 있도록 지원합니다.

Gemini Enterprise Agent Platform – Security and Governance

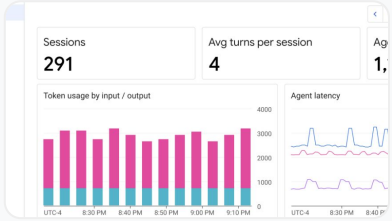
Gemini Enterprise Agent Platform

Build agents faster



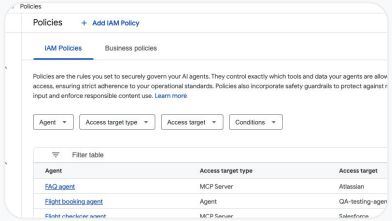
Create sophisticated agents using 200+ models, our upgraded ADK, and the new Agent Studio.

Scale agents effectively



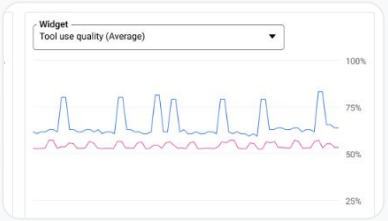
Power high-performance execution with Agent Runtime and long-term context via Memory Bank.

Govern agents with confidence

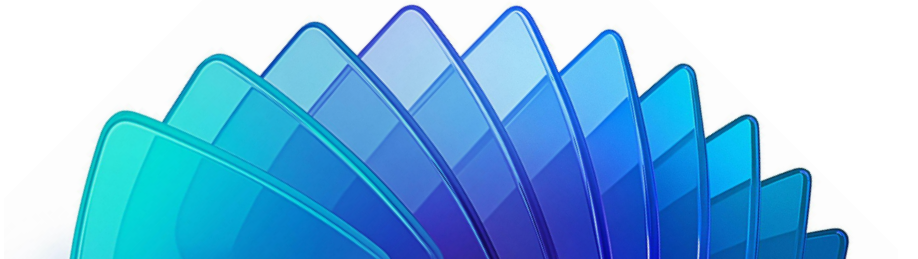


Establish centralized, zero-trust control using Agent Identity, Agent Registry, and Agent Gateway.

Optimize agents



Guarantee quality and trace reasoning loops with Agent Simulation, Evaluation, and Observability.



Gemini Enterprise Agent Platform

Build

- Agent Development Kit (New)
- 3P agent frameworks
- Agent Studio
- Agent Garden

Gemini API and Model Garden

- Gemini models
- 3P and open models
- Model training
- Model inference

Tools, data, and other agents

- A2A
- Grounding
- RAG
- MCP
- Search
- APIs and connectors
- A2UI
- AP2 and UCP
- Cloud Marketplace

Scale

- Agent Runtime (GA)
- Agent Sessions (GA)
- Agent Sandbox (GA)
- Agent Memory Bank (GA)

Govern

- Agent Gateway (New)
- Agent Identity (GA)
- Agent Registry (New)
- Agent Anomaly Detection (New)
- Model Armor
- Agent Policy
- Agent Security (New)
- Agent Compliance

Optimize

- Agent Evaluation (New)
- Agent Simulation (New)
- Agent Observability (New)
- Agent Optimizer (New)

AI Live + Labs

Preview

Agent Registry

Centralized catalog

Single source of truth for every AI agent, tool, and MCP server

Supports your Agents everywhere

Registry

에이전트 및 도구의 중앙 집중식 탐색과 확장 가능한 거버넌스

Centralized Agent Registry

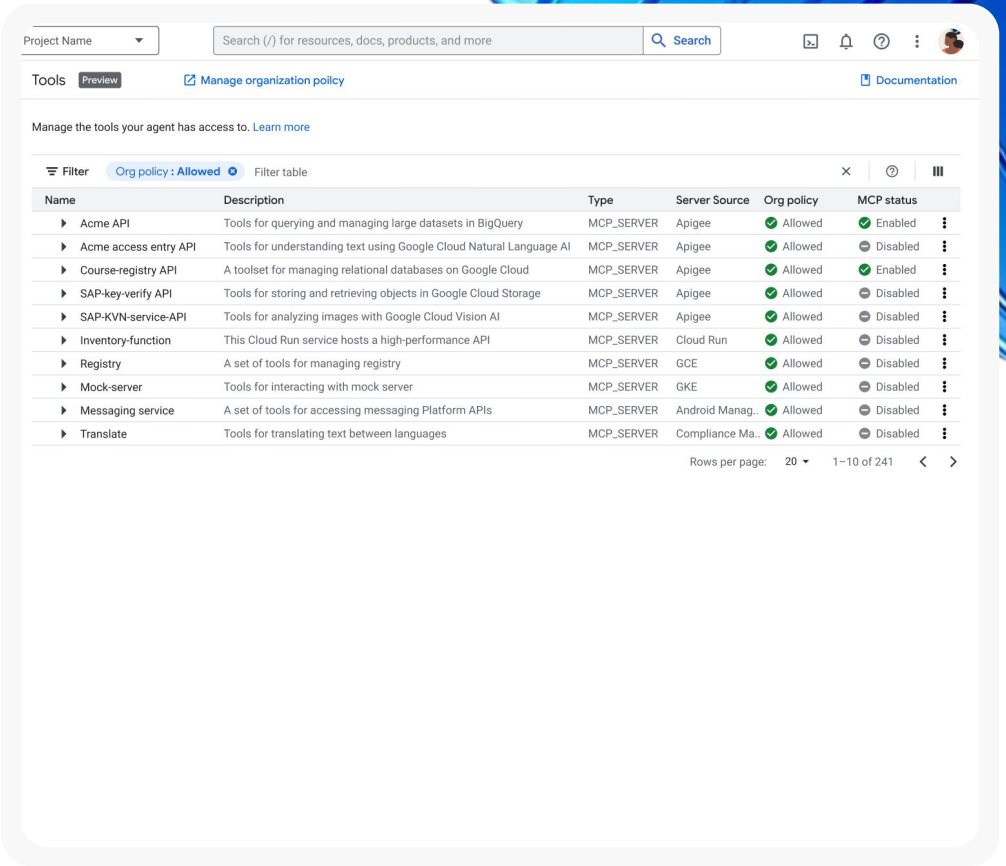
다양한 환경 전반에서 에이전트 개체에 대한 확장 가능한 거버넌스 및 관리를 제공합니다.

Tool Registry and Observability

도구 성능과 기업 내 사용 패턴에 대한 심층적인 인사이트를 바탕으로 기능의 중앙 집중식 탐색 및 재사용을 지원합니다.

Direct integration with policy

등록 상태, 특정 도구의 특성, 그리고 컴플라이언스와 같은 맞춤형 메타데이터 태그를 기반으로 접근을 제한하여 정밀한 거버넌스를 실행합니다.



GA

Agent Identity

제로 트러스트 보안의 핵심:
사람·시스템과 동등한 독립 주체로 인정

모든 에이전트에 위·변조 불가능한 고유 ID 자동 발급

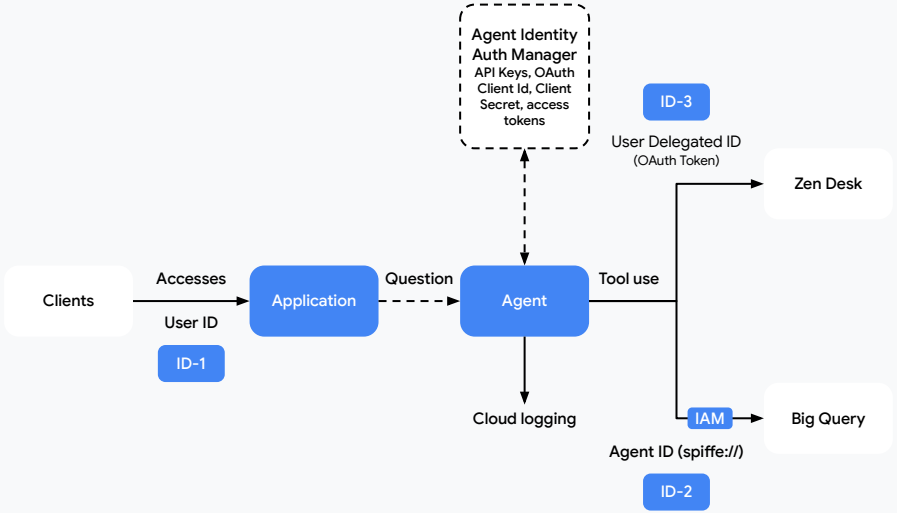
전체 에이전트 군단이 통용하는 만능 여권

Identities in Agentic Apps

Identity type	Purpose	Issuing system
ID-1 User Identity	사용자가 에이전트 또는 SaaS(예: Gemini Enterprise)에 직접 접근하기 위해 사용하는 신원	사용자가 가입할 때 인증을 처리하는 인간 IdP (예: Entra, Cloud Identity, Auth0 등)
(ID-2) Agent Identity (Using its own authority)	에이전트가 자체적인 고유 권한을 기반으로 백엔드 리소스에 접근할 때 사용하는 신원	에이전트가 생성될 때 발급하는 클라우드 플랫폼 (GCP)
(ID-3) 사용자 위임 에이전트 신원	에이전트가 사용자의 권한을 위임받아 사용자 맞춤형 리소스에 접근할 때 사용하는 신원	사용자가 OAuth 인증 절차(OAuth Dance)를 통해 접근을 위임할 때 발급하는 OAuth 서버 (1P/3P)

Example Scenario

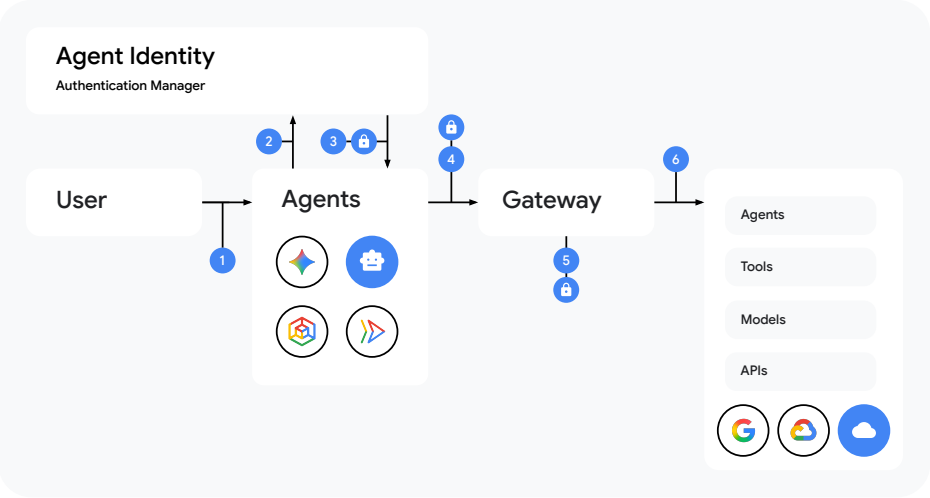
사용자가 본인 계정(ID-1)으로 고객지원 앱에 로그인한 뒤 질문을 남기면, 에이전트는 자체 권한(ID-2)으로 BQ 데이터를 가져오고 사용자 위임 권한(ID-3)으로 Zendesk 데이터를 조회해 최적의 답변을 제공합니다.



Managed Authentication for Agents

Agent Identity Auth Manager를 통한 최종 사용자 권한 위임 지원

Agent Identity Auth Manager는 최종 사용자의 토큰을 획득하고 안전하게 보관하는 금고(Vault) 역할을 수행합니다. Agent Gateway와 함께 사용할 경우, 모든 OAuth 토큰은 완전 암호화되어 에이전트 자체에는 노출되지 않습니다.



1. **User Delegation:** 사용자가 에이전트와 상호작용하며, 특정 도구나 다른 에이전트를 호출해야 할 때 필요한 접근 권한을 에이전트에 위임합니다.
2. **Credential Request:** 에이전트는 대상 도구 실행에 필요한 자격증명을 Auth Manager에 요청합니다.
3. **Token Issuance:** 보안 금고 역할을 하는 Auth Manager가 토큰(게이트웨이 미사용 시) 또는 암호화된 토큰(게이트웨이 연동 시)을 발급합니다.
4. **Secure Handoff:** 에이전트는 암호화된 토큰을 첨부하여 도구 실행 요청을 Agent Gateway로 전송합니다.
5. **Token Decryption (optional):** Gateway가 요청을 가로채 토큰을 복호화하고, 정의된 인가(AuthZ) 정책을 적용 및 검증합니다.
6. **Authenticated Execution:** 검증된 토큰이 최종 대상 도구로 전달되어 안전하게 실행됩니다.

Authz Policy

Controls to enforce intent-aware policies for Agent access

GCP IAM policy

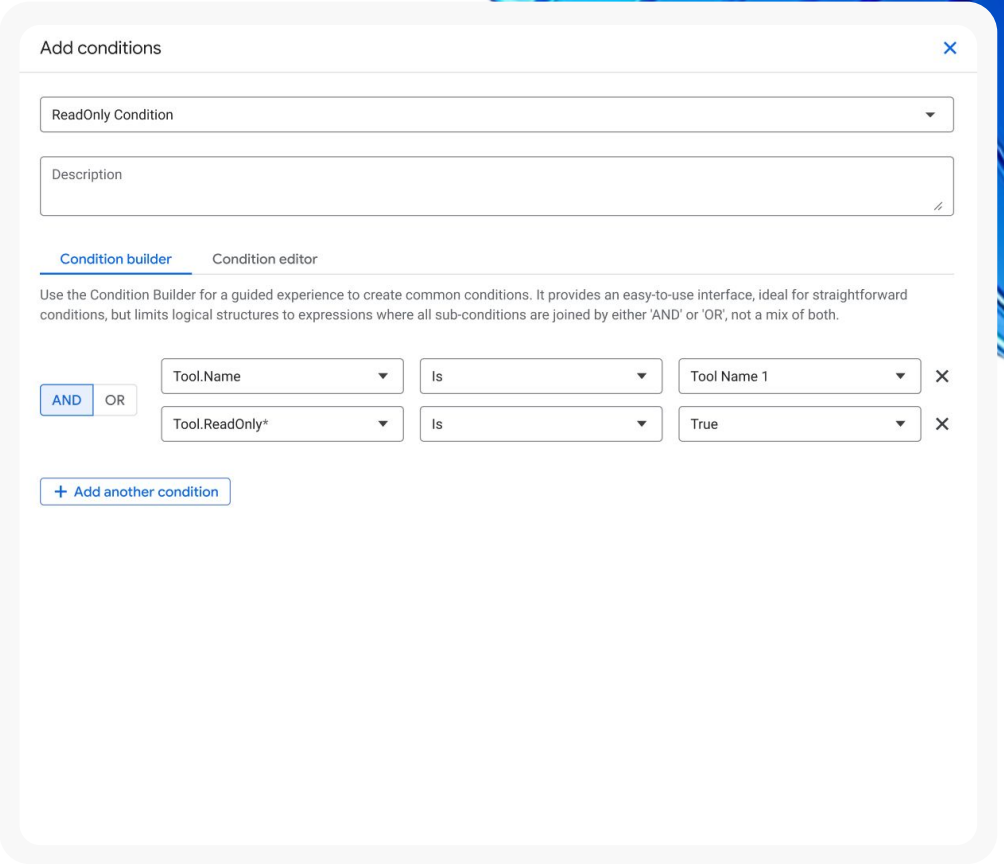
관리자는 MCP(Model Context Protocol) 서버에 접근 가능한 공인된 에이전트 및 클라이언트를 정의하는 GCP IAM 정책을 생성합니다. 등록되지 않은 클라이언트 및 에이전트의 접근은 원천 차단됩니다.

Dynamic Policy Engine

자연어 제약 조건(Conseca 기반)을 통해 실시간 컨텍스트에 맞춰 비즈니스 규칙을 동적으로 평가하며, 아웃오브밴드(Out-of-band) 방식으로 정책을 강제합니다.

Extensible Policy Fabric

오픈소스 표준 Service Extensions를 통해 고객 및 ISV(독립 소프트웨어 벤더)가 에이전트 데이터 플레인(Data Plane)에 맞춤형 거버넌스 로직을 직접 주입할 수 있습니다.



Preview

Agent Gateway

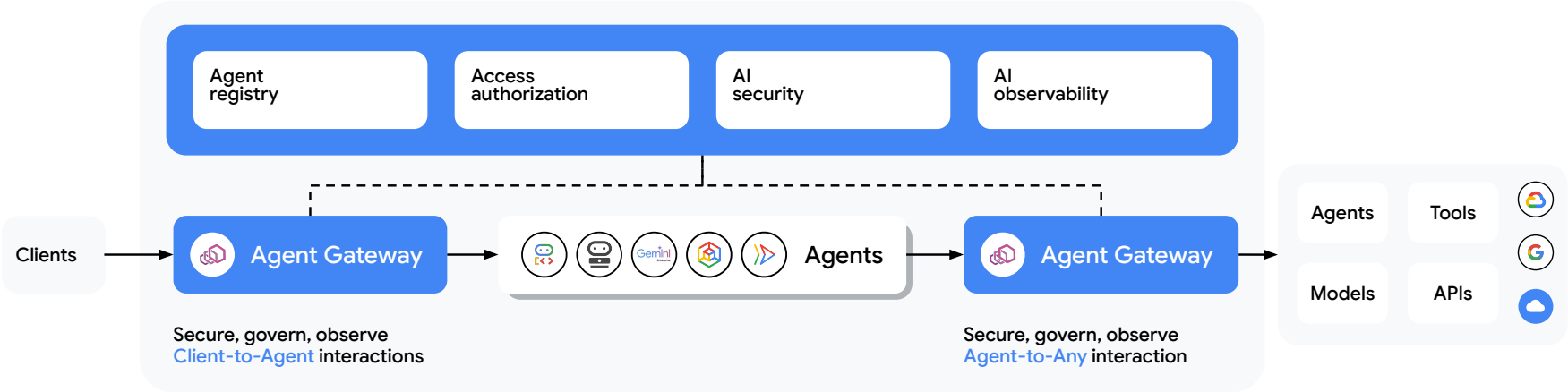
전사 AI 에이전트 통합 관제 체계 (“관제탑”) 구축

이기종 런타임 전반의 일관된 보안·식별(Identity)·거버넌스 체계 확립

오픈 스탠다드(Open Standards) 기반 아키텍처 구현

Agent Gateway

Consistent security and governance for every agent call



Service extensions

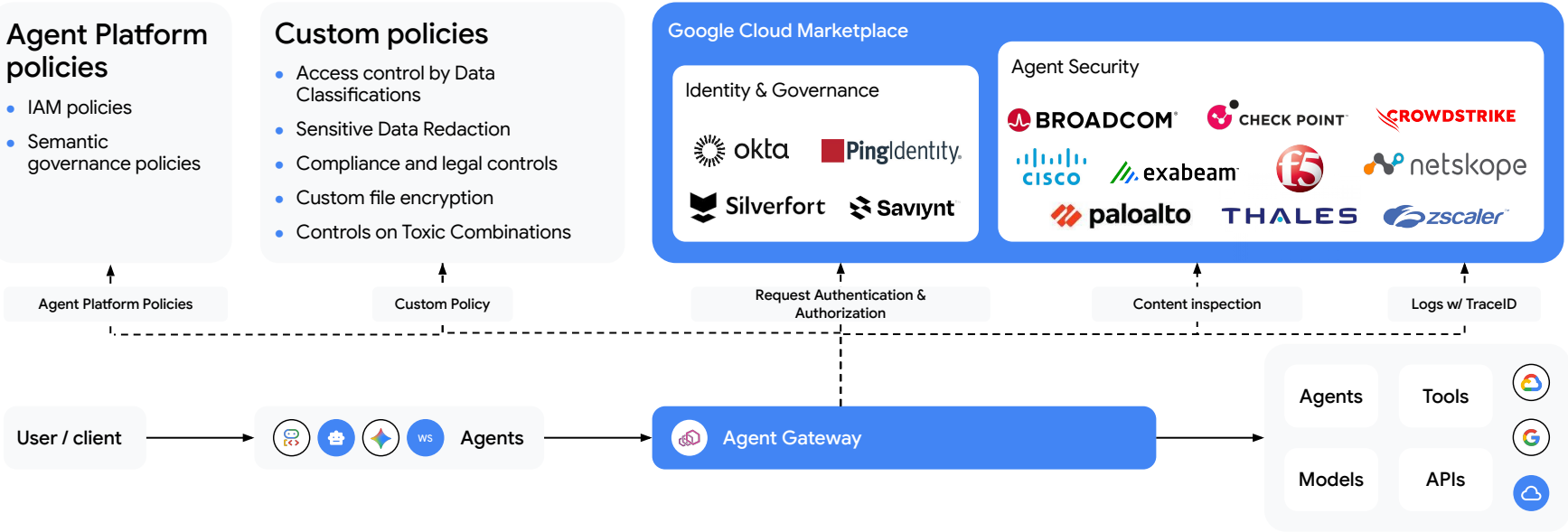
Providing policy extensibility for Agent Gateway in every path

Agent Platform policies

- IAM policies
- Semantic governance policies

Custom policies

- Access control by Data Classifications
- Sensitive Data Redaction
- Compliance and legal controls
- Custom file encryption
- Controls on Toxic Combinations



AI Live + Labs

Preview

Agent Security

Intelligent, automated systems to continuously monitor, detect, and alert you to suspicious activity and potential vulnerabilities

Agent Observability

Agent-specific telemetry

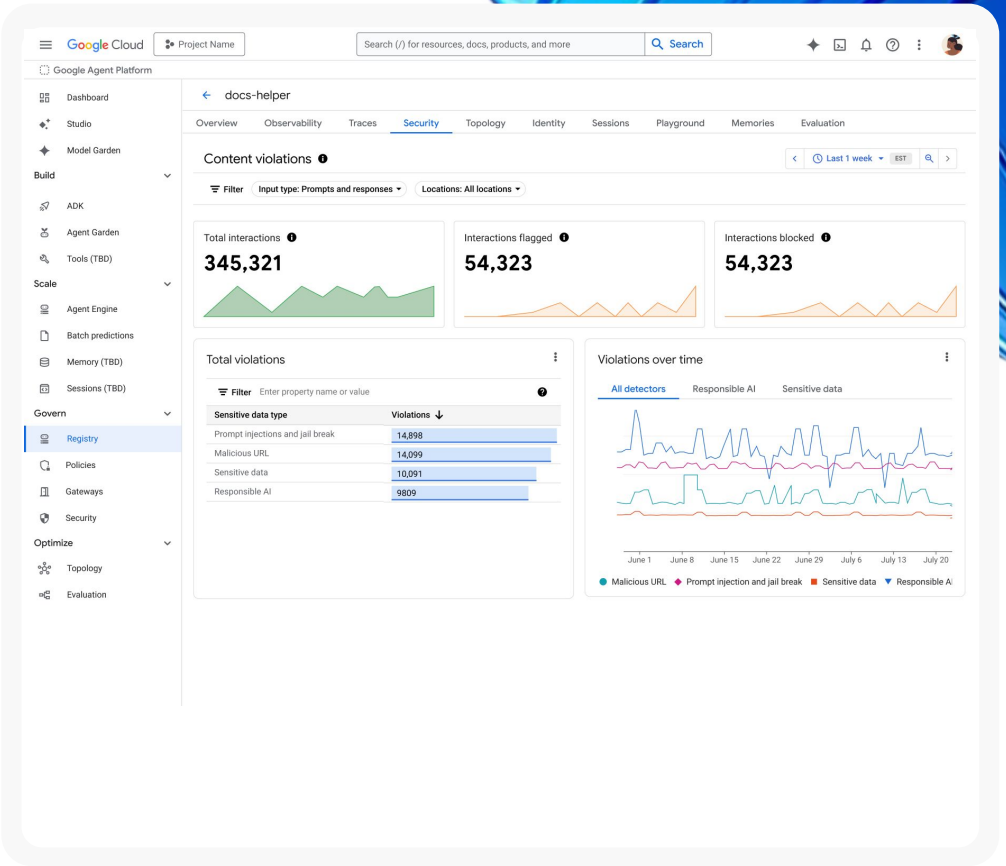
ADK에서 AI 특화 추적 데이터(Traces), 토큰 사용량, 도구 호출 내역 및 프롬프트-응답 로그를 원클릭으로 수집합니다.

Auditability

에이전트 게이트웨이(Agent Gateway) 로그에 에이전트 신원(Identity) 정보가 결합되어, 에이전트 접근 현황에 대한 완전한 가시성을 제공합니다.

Agent and tool topology

에이전트 간(A2A) 및 에이전트-도구(A2T) 그래프를 통해 에이전트, 도구, 데이터 간의 상관관계를 시각화합니다.



Evaluations

Simple, scalable agent evaluations

Fully managed eval harness

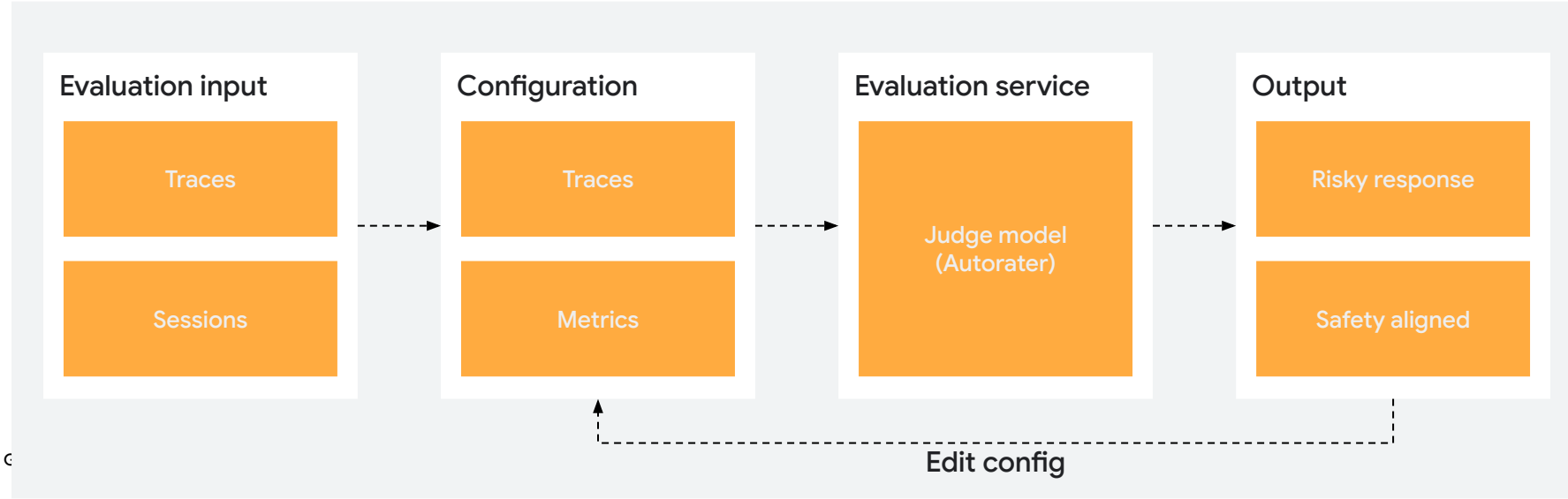
Agent Platform의 기본(Native) 기능으로 안전성(Safety), 품질(Quality), 보안(Security) 평가를 손쉽게 실행합니다.

Test task and safety adherence

에이전트가 민감한 데이터를 반환하거나 책임감 없는(Non-responsible) 방식으로 응답하는지 여부를 사전에 탐지하고

Session and trace based evals

이전 세션 이력이나 트레이스(Trace, 실행 추적 로그)를 평가하여 개발 환경과 프로덕션 환경의 성능 및 동작을 비교·분석합니다.



AI Live + Labs

Security Command Center

Detect vulnerabilities and active threats

Agent inventory

에이전트가 속한 프로젝트, 소유자, 생성 일시(타임스탬프), 배포 환경 등 에이전트 전반에 대한 상세 정보를 제공합니다.

Threat detection

의심스러운 활동을 실시간으로 모니터링합니다.

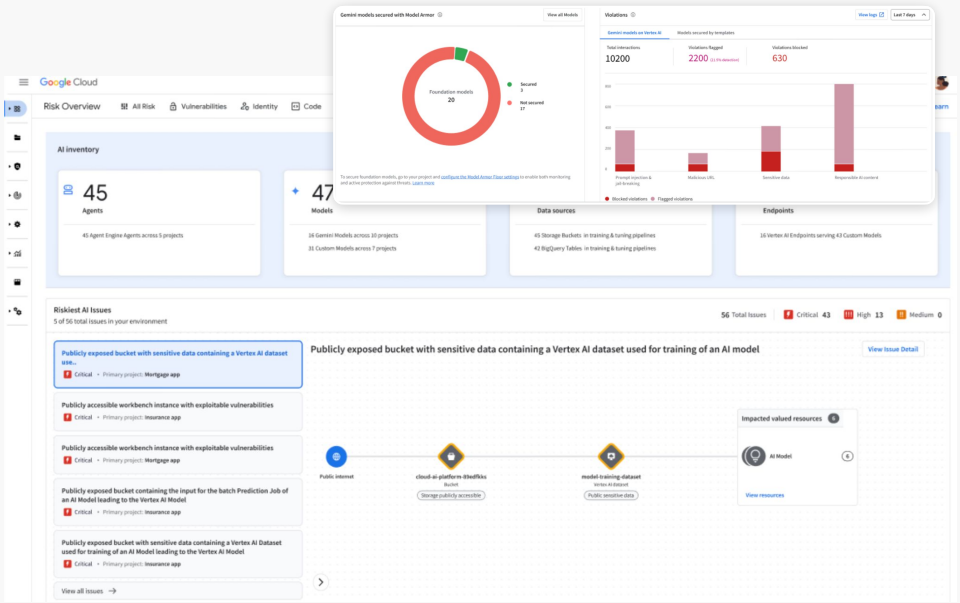
(예: 런타임 인스턴스 내 악성 코드 실행 등)

Agent vulnerability management

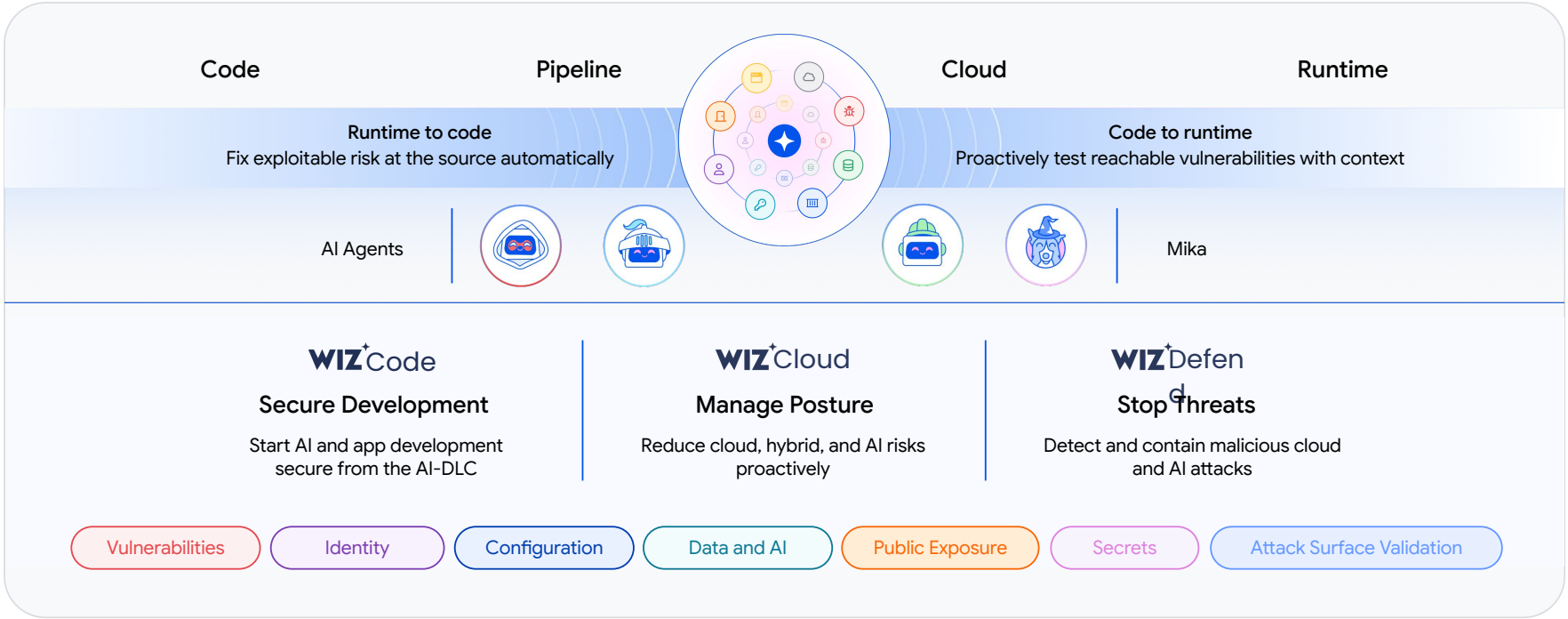
에이전트 자체 및 에이전트에 연동된 스킬(Skills/도구)의 보안 취약점을 사전에 탐지합니다.

Least privilege IAM recommendations

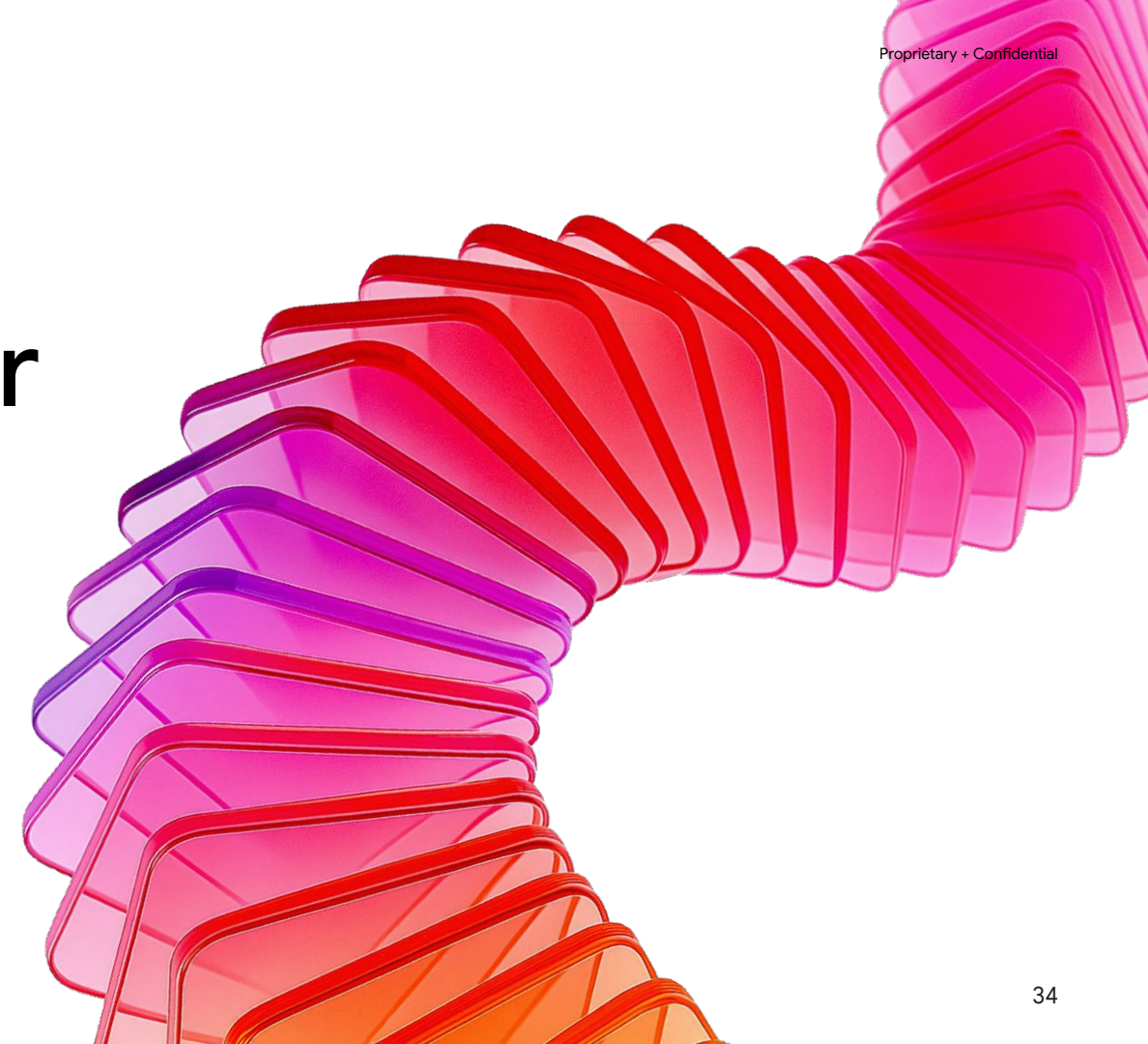
과도한 권한(Overprivileged)이 부여되었거나 장기간 사용되지 않는 에이전트 신원(Identity)을 식별하고, AI 기반의 최소 권한 최적화 가이드를 추천합니다.



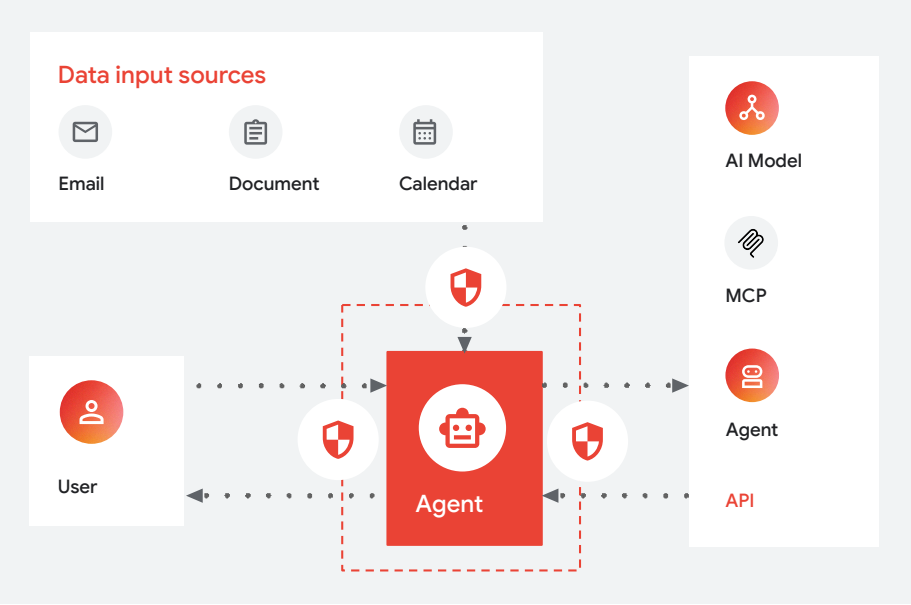
Wiz AI Application Protection Platform



Model Armor



Model Armor는 에이전트의 모든 상호작용 경로에 걸쳐 포괄적인 보호를 제공합니다



인라인 통합 (애플리케이션 수정 불필요)

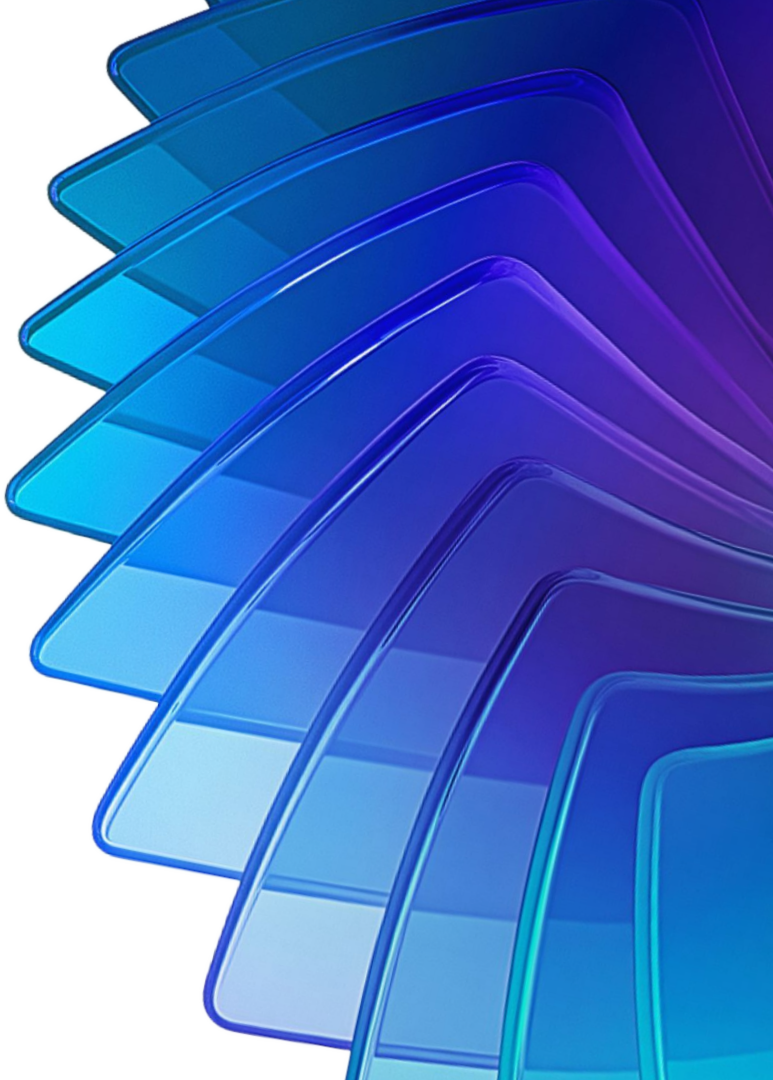
- • 신규 지원: Agent Gateway, Langchain, Firebase
- • 기존 지원: Gemini Enterprise, Agent Platform Runtime, Google MCP, Apigee, GKE-IG, LB
- • 커스텀 연동을 위한 API 제공

Protect user/model/agent interactions

- • 직접 및 간접 프롬프트 인젝션 차단
- • 위험하거나 공격적이고 유해한 콘텐츠 차단
- • 개인식별정보(PII) 비식별화(마스킹) 또는 차단
- • 악성 파일, 문서 및 안전하지 않은 URL 차단

AI Live + Labs

Thank you!



Google Cloud