AZURE DRAGON CYBER | WHITE PAPER

Secure AI Access to Customer Financial Data

A reference architecture for identity, AI agents and MCP, APIs, and SOC 2 controls in SaaS platforms

Mark Adams
Security architecture, GRC, and AI security governance
September 2026

Executive summary

Customers increasingly expect to ask an AI assistant about their own accounts: balances, holdings, fees, and recent activity. Delivering that experience safely is harder than securing a traditional web application, because an AI agent sits between the customer and the data and can be manipulated by the very text it processes.

This paper presents a vendor-neutral reference architecture, shown on AWS, for letting an external user authenticate to a SaaS application and retrieve their financial data through an AI assistant that calls APIs using the Model Context Protocol (MCP). It identifies 16 control points: ten technical controls along the request path and six SOC 2 program controls that make the design auditable over time.

Three ideas carry the design:

- **Identity follows the user.** The agent always acts on the customer's behalf with a narrowed, user-bound token, never with broad credentials of its own.
- **One policy decision, enforced in layers.** A central policy engine decides; the API gateway, MCP layer, and financial API each enforce the result.
- **Every control produces evidence.** Controls are mapped to SOC 2 Trust Services Criteria, NIST SP 800-53, and the NIST AI RMF so one control set supports all three.

Design goal: the AI never sees more than the customer could see directly, and every control leaves evidence an auditor can test.

Why AI changes the threat model

A conventional portal has a predictable path: the user clicks, the app calls an API, the API returns the user's records. Adding an AI assistant introduces a component that interprets natural language, chooses which tools to call, and composes answers from retrieved data. That creates risks that traditional application controls do not fully address:

- **Prompt injection.** Instructions hidden in user input or retrieved content try to make the agent fetch other customers' data or ignore its rules.
- **Confused deputy.** An agent holding broad service credentials can be tricked into using them for a request the user was never entitled to make.
- **Broken object-level authorization.** APIs that trust an account ID supplied by the agent can leak another customer's records.
- **Token theft and passthrough.** Long-lived or reused tokens let an attacker, or a compromised tool, replay a customer's access.
- **Data leakage in outputs.** Answers may expose account numbers, personal data, or content the user should not see.
- **Unverifiable behavior.** Without end-to-end logging, no one can reconstruct what the agent did, which defeats incident response and audit.

NIST's AI Risk Management Framework describes the outcomes an organization should achieve for AI systems, and its Generative AI Profile adds suggested actions for risks such as information security and data privacy.[1][2] Neither prescribes a specific architecture. NIST is developing SP 800-53 control overlays for AI systems, including single-agent and multi-agent use cases, but they are not yet final.[3] This paper fills that gap with a concrete design.

1. NIST, AI Risk Management Framework (AI RMF 1.0), https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf
2. NIST, Generative AI Profile (NIST AI 600-1), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
3. NIST CSRC, SP 800-53 Control Overlays for Securing AI Systems (COSAiS), https://csrc.nist.gov/projects/cosais

Five design principles

Principle	What it means in practice
1. Identity follows the user	Every agent and tool call carries a short-lived token bound to the customer and scoped to the task. The agent has no standing access to customer data.
2. One policy, enforced in layers	Authorization lives in a central, deny-by-default policy engine. The gateway, MCP layer, and API all enforce it, so bypassing one layer is not enough.
3. AI is untrusted input and output	Guardrails screen prompts and responses. The assistant is read-only by default; any action that moves money needs explicit human approval.
4. Isolation down to the keys	Tenant separation is enforced by row-level security and per-tenant KMS encryption context, not only by application logic.
5. Every step is evidence	A single correlation ID links prompt, tool call, policy decision, API call, and data access for forensics and audit sampling.

Authentication and delegation flow

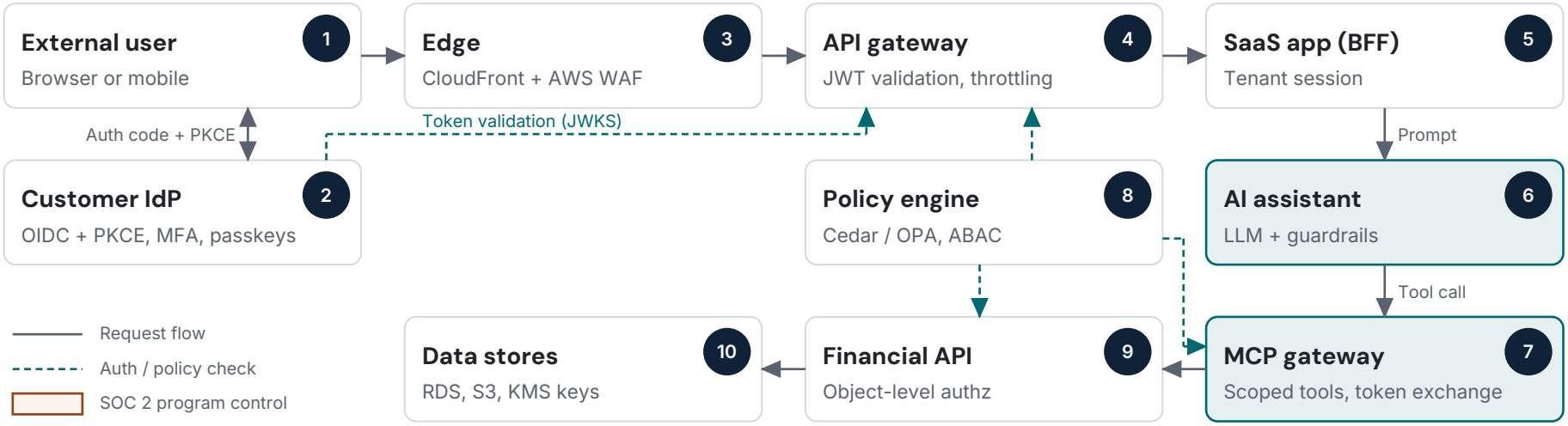
Authentication starts with the customer, not the workforce directory. Customers use a customer identity platform such as Amazon Cognito, Microsoft Entra External ID, or Okta Customer Identity. The flow then carries that identity through every hop:

Step	What happens	Key control
1. Sign in	The browser is redirected to the customer IdP, which authenticates with passkeys or FIDO2 MFA and returns an authorization code.	OIDC authorization code + PKCE; risk-based step-up
2. Session	The backend for frontend (BFF) exchanges the code for tokens and stores them server-side. The browser receives only a session cookie.	HttpOnly, Secure, SameSite cookies; no tokens in the browser
3. Validate	The API gateway validates each request's token against the IdP's published keys.	Signature, issuer, audience, expiry, and scope checks
4. Delegate	When the assistant calls a tool, the MCP gateway exchanges the user token for a short-lived token scoped to that tool and bound to the user.	OAuth 2.0 token exchange (RFC 8693); no token passthrough
5. Authorize	The financial API asks the policy engine whether this user owns the account and has consented, then returns only required fields.	Object-level authorization; data minimization

Why step 4 matters: even if a prompt injection convinces the agent to request another customer's data, the delegated token and the policy engine both refuse, because the agent can never exceed the customer's own entitlements.

Reference architecture

Numbered badges mark control points. Points 1 to 10 are technical controls on the request path; points 11 to 16 are SOC 2 program controls. Dashed teal lines are authentication and authorization checks.



SOC 2 program controls (Trust Services Criteria)



Technical control points

Each control point lists the core technical controls, the SOC 2 Trust Services Criteria it supports, and the evidence an auditor is likely to sample.

#	Layer	Key technical controls	SOC 2	Audit evidence
1	External user	Passkey enrollment, device and bot signals, no credentials or account numbers typed into chat, AI-use disclosure	CC2.3, P1, P2	Privacy notice, terms, consent records, AI disclosure text
2	Customer IdP	OIDC + PKCE, phishing-resistant MFA, risk-based step-up, secure recovery, short-lived tokens with refresh rotation, identity proofing	CC6.1, CC6.2	MFA and token-lifetime configuration, recovery procedure, sample of registrations
3	Edge	TLS 1.2+ and HSTS, WAF managed rules for OWASP Top 10, bot control, rate limits, DDoS protection	CC6.6, CC6.7	WAF rule set, TLS scan results, blocked-attack reports
4	API gateway	Token signature, issuer, audience, expiry and scope checks; per-tenant throttling; schema validation; private backends only	CC6.1, CC6.6, A1.1	Authorizer configuration, API inventory, throttling settings
5	SaaS app (BFF)	Server-side tokens, HttpOnly cookies, CSRF protection, tenant ID taken from the token, idle and absolute timeouts	CC6.1, CC7.1	Penetration test and remediation, SAST/DAST results, session configuration
6	AI assistant	Private model endpoint, no training on customer data, prompt-injection and PII guardrails, grounded answers, versioned prompts, read-only default	CC3.2, CC7.2, PI1.3, C1.1	AI risk assessment, guardrail configuration, evaluation and red-team results
7	MCP gateway	OAuth-based tool authorization, token exchange to user-bound scoped tokens, no passthrough, tool allowlist, per-tool scopes, human approval for writes	CC6.1, CC6.3, CC7.2	Tool inventory with scopes, approval logs, tool registry change history
8	Policy engine	Deny by default; attribute-based rules on user, tenant, account ownership and consent; every decision logged	CC5.2, CC6.1, CC6.3	Version-controlled policies, policy test cases, decision logs
9	Financial API	Object-level authorization, consent check, field minimization, masked account numbers, mTLS or IAM for service calls	CC6.1, PI1.2-PI1.4, P6	Authorization test results, consent records, data accuracy reconciliation
10	Data stores	KMS customer-managed keys with tenant context, row-level security, private subnets, VPC endpoints, immutable backups	CC6.1, C1.1, C1.2, P4, A1.2	Key policy and rotation, retention schedule, disposal records, restore test

SOC 2 program controls

SOC 2 examinations evaluate an organization's controls against the AICPA's 2017 Trust Services Criteria, with points of focus revised in 2022.[4] Architecture alone does not satisfy them. These six program controls provide the operating discipline around the design:

4. AICPA, 2017 Trust Services Criteria (Revised Points of Focus, 2022), https://www.aicpa-cima.com/resources/download/2017-trust-services-criteria-with-revised-points-of-focus-2022

#	Program control	Criteria	What to implement
11	Governance and risk	CC1-CC3, CC5.3	Board or committee oversight, code of conduct, background checks, security and AI training, annual risk assessment including fraud and AI risks, approved policies.
12	Change management	CC8.1	Peer review, CI/CD approvals, separation of developers and deployers. Treat model, prompt, guardrail, and MCP tool changes as formal changes.
13	Vendor and subservice risk	CC9.2	Inventory the cloud provider, model provider, IdP, and any data aggregator; review their SOC reports annually; contract for no-training and breach notification terms.
14	Availability and resilience	A1.1-A1.3, CC9.1	Multi-AZ design, capacity monitoring, backups, defined RTO and RPO, tested business continuity and disaster recovery plans.
15	Monitoring and incident response	CC4.1, CC7.1-CC7.5	Centralized logging and SIEM alerting, vulnerability scanning, correlation IDs across prompt, tool and API, AI-specific incident playbooks.
16	Workforce and privileged access	CC6.1-CC6.3, CC6.8	Joiner-mover-leaver process, SSO and MFA for staff, privileged access management with just-in-time admin, quarterly access reviews, endpoint protection.

What SOC 2 means for AI systems

- **Operating effectiveness over time.** A Type II report tests whether controls worked across the audit period, typically six to twelve months. Access reviews, change approvals, alert handling, and DR tests need dated evidence each time they occur.
- **Subservice organizations.** The cloud provider, model provider, and IdP are subservice organizations. Their controls are either excluded from scope and referenced (carve-out) or included and tested (inclusive).[5]
- **Complementary user entity controls.** The system description lists controls customers must perform themselves,[6] such as protecting their own devices and credentials and, for business tenants, managing their own users.

- **No AI-specific criteria yet.** The AICPA has not published AI-specific Trust Services Criteria, so auditors are applying existing criteria to AI.[7] Map AI risk assessment to CC3.2, prompt, model and tool changes to CC8.1, the model provider to CC9.2, and the accuracy of AI answers about financial data to processing integrity (PI1).

5. Securisea, SOC 2 Carve-Out Method vs Inclusive Method, https://www.securisea.com/resources/soc-2-carve-out-method-vs-inclusive-method

6. Linford & Company, SOC 2 Description Criteria, https://linfordco.com/blog/what-are-soc-2-description-criteria/

7. Cloud Security Alliance, The SOC 2 AI Gap, https://labs.cloudsecurityalliance.org/research/csa-research-note-soc2-ai-controls-gap-20260830-csa-styled/

One control set, three frameworks

The same controls can support SOC 2, NIST SP 800-53 Rev. 5, and the NIST AI RMF. The mapping below reflects practitioner judgment; no official crosswalk links all three. For a broader AI-specific control catalog, the Cloud Security Alliance AI Controls Matrix maps its controls to the NIST AI RMF, AI 600-1, ISO/IEC 42001, and the EU AI Act.[8]

Control area	SOC 2	NIST SP 800-53 Rev. 5	NIST AI RMF
Customer and agent identity	CC6.1, CC6.2	IA-2, IA-5, IA-9	Govern, Manage
Authorization and tenant isolation	CC6.1, CC6.3	AC-3, AC-4, AC-6	Manage
Prompt and output guardrails	CC7.2, PI1.3	SI-10, SI-15	Measure, Manage
Model, prompt and tool change control	CC8.1	CM-3, CM-4	Govern, Manage
AI vendor and model provider	CC9.2	SA-9, SR-5, SR-6	Govern 6
Encryption and data protection	CC6.1, C1.1, C1.2	SC-8, SC-12, SC-28	Manage
Logging, monitoring, incident response	CC7.2-CC7.4	AU-2, AU-12, SI-4, IR-4	Measure, Manage 4
Testing and AI red teaming	CC4.1, CC7.1	CA-8, RA-5, SA-11	Measure 2

8. Cloud Security Alliance, AI Controls Matrix v1.1, https://cloudsecurityalliance.org/blog/2026/07/14/ai-controls-matrix-v1-1-strengthening-the-foundation-for-trustworthy-ai

Implementation roadmap

Phase	Focus	Outcomes
1. Foundation	Identity and authorization	Customer IdP with phishing-resistant MFA, BFF session pattern, API gateway token validation, central policy engine, object-level authorization tests.
2. Secure AI	Assistant and MCP layer	Private model endpoint, guardrails, read-only tool allowlist, token exchange at the MCP gateway, human approval for writes, AI red-team baseline.
3. Assurance	SOC 2 readiness	Control owners and evidence standards, AI changes in change management, vendor reviews, correlation-ID logging, DR test, readiness assessment before the Type II period.

Conclusion

AI assistants can give customers faster, more natural access to their financial information, but only if the architecture assumes the agent can be manipulated. Binding every call to the customer's identity, centralizing authorization, treating AI input and output as untrusted, and building SOC 2 evidence into daily operations turns a risky integration into a defensible, auditable service.

About the author

Mark Adams is a senior IT risk, cybersecurity, and GRC professional with more than 30 years in technology, spanning security architecture, cloud and identity security, audit and assurance, incident response, and AI security governance. Through Azure Dragon Cyber he provides cybersecurity training, mentoring, and consulting, including NIST CSF 2.0, NIST AI RMF, and SOC 2 readiness assessments.

This paper is a reference design for educational purposes. It is not legal or audit advice, and controls should be tailored to each organization's risks, regulators, and auditors.