Redpanda

2026

Agentic Governance Reference Architecture

Govern AI agents in production, on any stack

We built governance for people. Not agents.

For a century, organizations have managed imperfect people with infrastructure, not trust. Identity, access controls, activity monitoring, escalation paths, the authority to revoke—these mechanisms are so ordinary they're nearly invisible. None of them assumed good character. What they did assume was a participant with a body, a history you could hold them to, and a working day.

AI agents break all of those assumptions. They move at machine tempo, faster than any approval process built for human review. They operate at machine scale, where one delegated task fans out into hundreds of processes touching dozens of systems. They're compliant to a fault, yet leave nothing behind to hold them accountable. Alignment training installs behavior that is prompt-conditional, not value-grounded, so what you get is a convincing imitation of judgment, which breaks down when tested.

The mimic is not the thing mimicked. And this disconnect is what keeps agents stuck in pilots.

So the question isn't "Can agents run in production?", it's "Can you trust them when they do?" That takes **governance infrastructure**, not good intentions. One place to set policy and a record of every action an agent took. One switch to stop a rogue agent before it causes damage. And all of it must run on open standards that you control, not a single vendor's walled garden, so it works with the systems you already have and the ones you might adopt in the future.

With that question in mind, this paper presents the foundation for the answer. First, an introduction to the Agentic Data Plane and the use cases it enables enterprises to run in production. Then, a vendor-agnostic reference architecture for governing agents (a pattern you could build on any stack), and how Redpanda implements it end-to-end.

Head of AI | Chief Technology Officer, Redpanda

The gap in governance infrastructure

AI agents are arriving in the enterprise, where the infrastructure to govern them doesn't exist yet. These agents can access your systems, pull records, act on behalf of real people, and spend real money against external models. Most tech stacks weren't built for any of this. Service accounts can't carry user identity through a tool call, API gateways don't know how to enforce per-agent token budgets, and your SIEM can't reason about a chain of LLM steps and tool invocations.

The **Agentic Data Plane** fills that gap.

It's the governance layer between your AI agents and everything they touch. It gives every agent its own identity, brokers every tool call, enforces policy at runtime, and records the full trace of what each agent did and why. It's a single, central layer that clearly answers the questions production agents usually raise: Which agent did this? What was it allowed to touch? What did it actually do? What did it cost?

This layer enforces the four functions that directly address how existing infrastructure fails agents today: **identity, access, evidence, and control**.

Identity

Every agent acts as a known user, not a shared service account.

Access

Every tool call is brokered, scoped to least privilege per system.

Evidence

Every action is recorded: what the agent did, on whose behalf, and what it cost.

Control

Policy is enforced at runtime, and consequential actions wait for human approval.

Critically, there's one principle that everything depends on: all four must run through channels the agent cannot access, modify, or circumvent.

Any policy enforced through the agent is only as strong as the agent's ability to perfectly retain and obey it. Put rules in the system prompt, and prompt injection can override them entirely. Train the agent to respect boundaries, and hallucination can cause it to confidently invent permissions it doesn't have. Even routine context management can silently drop the rules it was told to follow. Real governance runs out-of-band, outside the agent's data path, invisible to it, and enforced by infrastructure the agent cannot touch.

Another key piece of this governance infrastructure is its portability.

Consider what your agents actually need to reach: your data isn't in one cloud, your SaaS isn't from one vendor, your models won't be from one provider this year or next. A walled-garden Agentic Data Plane can govern part of the workflow within its walls. The rest is invisible to it, which means it's invisible to you.

Open standards are what make a governance layer cover the ground it claims to cover. MCP gives you one protocol for agent-to-tool communication. OpenTelemetry gives you one trace format. OAuth and OIDC give you identity flows that work across vendors. Bet on these and your stack stays portable as the ecosystem moves underneath it.

Getting this infrastructure right is how you run and scale agents in production that you can trust. It's how you stop choosing between innovation and control.

> "The agents aren't the problem. The missing infrastructure between agents and your data is the problem."

Use cases

From pilot to production, safely

The block to getting agents into production isn't the model, it's governance. With the right infrastructure, the use cases below can finally move from demo to deployment.

Internal assistants

Agents over systems of record

Answer and act across CRM, ITSM, HR, and internal databases. Every agent runs as the signed-in user, with least-privilege access to each system.

Regulated data

Customer-facing agents on sensitive data

Support and servicing agents that touch Personally Identifiable Information (PII) or financial records, with policy enforced and every action captured on the record.

Autonomous workflows

Multi-step execution across systems

Reconciliation, claims, rebalancing. The agent plans and acts on its own, and consequential steps wait for human approval. A kill switch stops the rest.

Procured agents

Cross-vendor and bring-your-own agents

Vendor-sourced agents that arrive with their own identity models, but can be governed under the same enterprise policy, through the same plane as your own agents.

Whatever the framework, model, or cloud, the Agentic Data Plane works with the systems you already have. No rip-and-replace or walled gardens.

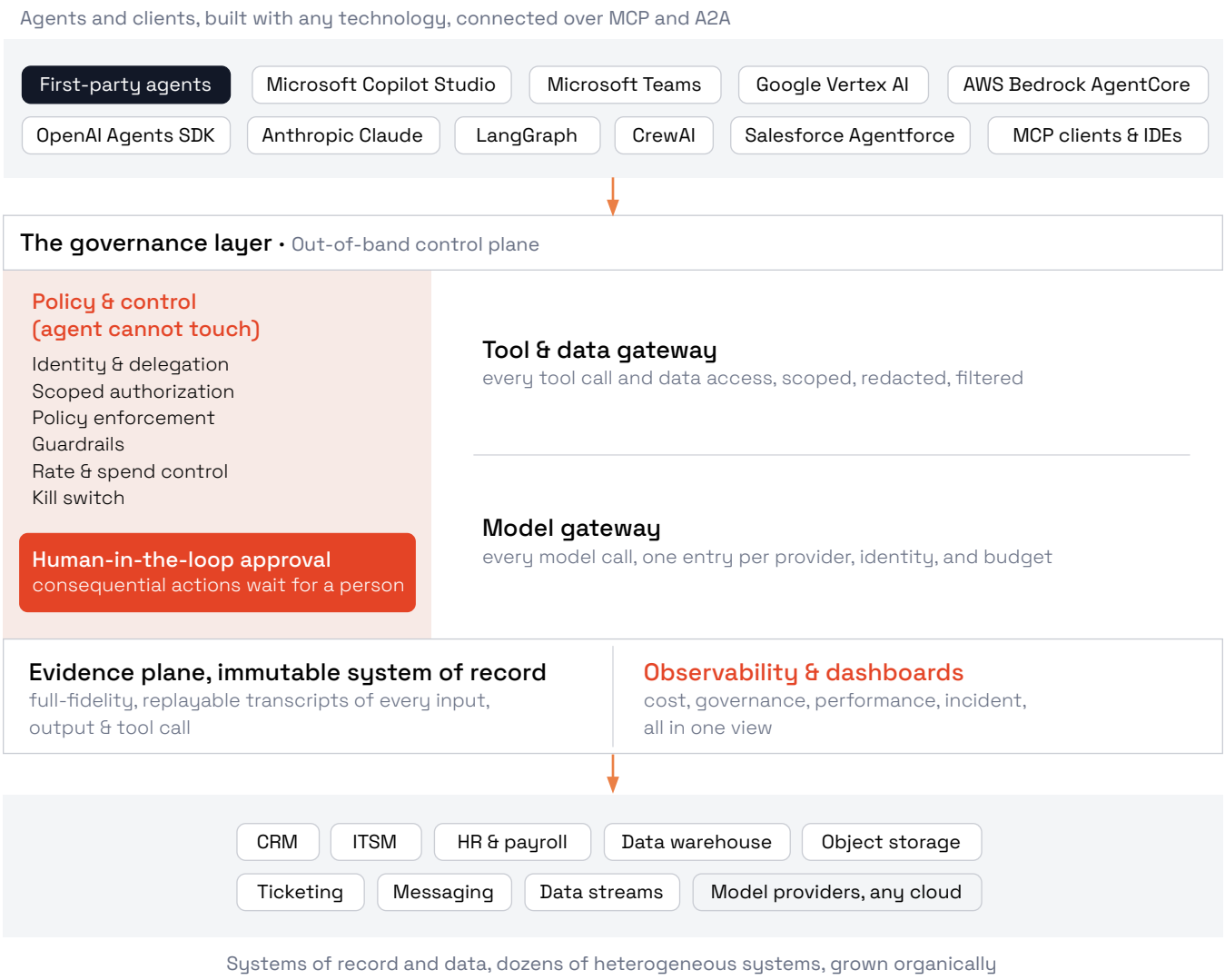
Observability, built in

Trust is earned, not given, which means you have to be able to see what every agent is doing. Bring your own models, your own agents, your own SaaS, your own IDP: the Agentic Data Plane governs them all the same way, in one place.

Cost	Governance	Performance	Incidents
Spend by agent, model, and team, with caps and budgets before the runaway bill.	Every identity, access, and policy decision, audit-ready and replayable.	Throughput, success and error rates, and latency for every agent in production.	Anomalies surfaced early, with a kill switch and full-transcript replay to resolve.

Reference architecture

A vendor-agnostic architecture for governing agents



The foundation: out-of-band control

One control layer sits between agents and the sprawl of systems they touch, enforcing identity, authorization, and policy through channels the agent cannot reach or modify. It keeps a human in the loop on consequential actions and turns every action into evidence. This gives you the freedom to build or buy agents and govern them all based on a centralized set of rules.

The architecture, in production today

The Redpanda Agentic Data Plane is a neutral governance layer for enterprise agents. It sits above the model and runs on open standards, so when the model, framework, or cloud changes, you don't rip anything out. It plugs into the systems you already run.

	Role in the architecture	Redpanda implementation
Identity	**Authentication**	Instance-bound, cryptographic identity for each agent, tied to a specific task and human
Access	**Tool & data gateway**	MCP Gateway (managed and self-managed servers)
	Model gateway	AI Gateway, one entry per provider
Evidence	**Evidence & observability**	OpenTelemetry transcripts on an immutable topic, queried with SQL Audit logs in OCSF standard format
Control	**Out-of-band control**	Per-user OAuth & token vault Deterministic authorization policies Semantic gating on sensitive actions Guardrails Spend caps Kill switch
	Human in the loop	Approval workflows on consequential actions
Open standards	**Interoperability & deployment**	MCP, OAuth, A2A, OpenTelemetry, Apache Kafka®. Runs in your own cloud (BYOC)

With a view for each owner

CFO

Spend by agent, model, and team, with caps before the runaway bill.

CISO

Who is acting as whom, what got blocked, and how to stop it.

GRC

Audit-ready policy decisions and full replayable transcripts.

CTO

Which agents are production-ready, which are failing, and why.

100+ agents built in the first week by design partner, a global semiconductor manufacturer in a regulated environment. Each one carried a per-user OAuth identity to its systems of record.

Learn more at redpanda.com/agentic-data-plane

What's next

Run agents on real data, safely

Most of the conversation about agent governance focuses on improving the models, tightening alignment, and reducing hallucinations. That work matters. But it's not what's stopping your agents from reaching production. What's missing isn't "perfect agents"; it's the institutional infrastructure that lets imperfect agents do real work **safely**.

An Agentic Data Plane is how you build it right, with identity, access, evidence, and the control to revoke an agent at any time.

The Redpanda Agentic Data Plane is built on infrastructure already trusted by Fortune 500 companies and fast-growing startups, with everything needed to run agents on real data under a centralized set of rules, so organizations never have to choose between moving fast and staying in control.

Want to see what governed agents look like in production?

Book a demo of the Redpanda Agentic Data Plane with one of our product experts.

Book a demo

Read the research

The new AI risk isn't what models say	Forbes.com ↗
Why your AI agents need performance reviews, too	Forbes.com ↗
Post-human: we all built agents, nobody built HR	O'Reilly.com ↗
Don't automate your moat	O'Reilly.com ↗
Governing AI agents in production	Redpanda.com ↗
Agents you can trust with centralized AI governance	Redpanda.com ↗
What is an Agentic Data Plane?	Redpanda.com ↗