

EBook

FinOps in Banking:

Building the Cloud+
Operating Model for AI, SaaS
& Technology Value



Table of Contents:

The Case For Change

Why AI Spend Is Breaking Traditional Banking Cost Models ----- 03

The Operating Model

From Cloud to Cloud+:
What Modern FinOps Means in Banking ----- 05

What Is FinOps Maturity? ----- 06

Introducing The Cross-Functional FinOps
Operating Model For BFSI Organizations ----- 07

Which Technical Features
Define The Cloud+ FinOps Operating Model? ----- 10

From Model To Practice

How Agentification & Agentic FinOps Work In BFSI ----- 12

How AI Is Applied To
Banking FinOps: Real-World Signals ----- 14

Regulation, Risk And Resilience:
FinOps Under The Supervisors’ Gaze ----- 15

Your Next Steps: Introducing Ciklum PRODIGY ----- 17

Why AI Spend Is Breaking Traditional Banking Cost Models

In 2026, Uber became a cautionary tale for AI spending. Four months into the year, they had [reportedly used their entire annual AI budget](#), with trade press pointing to heavy adoption of AI coding tools such as Claude Code and Cursor. It wasn't a banking story, but banks should recognize the pattern: when an AI tool becomes useful, adoption can spread quickly, and the bill can grow just as fast.

That's what makes AI spend different. A SaaS license is relatively predictable. A per-token API model isn't. A developer assistant, fraud-detection model or internal workflow agent can move from pilot to everyday use in weeks, with costs shaped by activity, model choice, vendor pricing and the way teams use the tool.

The problem is already showing up in enterprise cost management. [A 2025 AI cost governance benchmark](#) found that 85% of companies miss AI cost forecasts by more than 10%, with nearly 25% missing by more than 50%. For banks, that's not just a budget problem. It affects forecasting, resilience, outsourcing risk, model governance and executive accountability.



85%

**of companies miss AI
cost forecasts by more
than 10%**

Why AI Spend Is Breaking Traditional Banking Cost Models

Three pressures are now coming together. First, the scope of banking technology spend has expanded far beyond public cloud. A single customer journey might touch core systems, private cloud, public cloud, SaaS platforms, AI services and third-party data, each with its own owner, billing format and level of visibility.

Second, forecasting is becoming less reliable. Banks are used to modeling infrastructure, licenses and service contracts, but AI introduces more variable costs, especially where models are charged by use. A successful internal AI tool can become a material cost center before the next budget cycle catches up.

Third, regulation raises the stakes. Banks need to show control over outsourcing, operational resilience, vendor concentration, data residency and model risk. Cost visibility helps, but it's not enough on its own. Leaders need a clear line of sight from spend to ownership, value and control.

This is the gap Cloud+ FinOps is built to close.

A Cloud+ FinOps operating model gives banks one view of cost, usage, ownership, risk and value across cloud, SaaS, private infrastructure, data center and AI workloads. The goal isn't simply to reduce spend. It's to answer a harder question before spend scales: is this technology investment creating proportionate business value, with the right controls in place?



From Cloud to Cloud+: What Modern FinOps Means in Banking

The FinOps Foundation defines FinOps as

an operational framework and cultural practice that maximizes the business value of technology, enables timely data-driven decisions, and creates financial accountability through collaboration between engineering, finance and business teams.

For banks, the important shift is from cloud to cloud and technology.

FinOps was first formalized around public cloud, when the main challenge was understanding and optimizing spend across hyperscaler environments. But banking technology spend now reaches much further. A mortgage application, fraud-detection model or onboarding journey might draw on public cloud, private cloud, SaaS, data platforms, AI models and legacy systems.

A public-cloud-only FinOps practice can tell a bank what it spent with one provider last quarter. A Cloud+ FinOps practice can show what a product, channel, model or transaction costs across the full stack. That's a much more useful question, especially when senior leaders are trying to decide which technology investments are worth scaling.

It's also why the FinOps Foundation introduced Scopes into the framework, so FinOps practices could apply across different areas of technology usage and cost, including public cloud, SaaS, data center and private cloud.

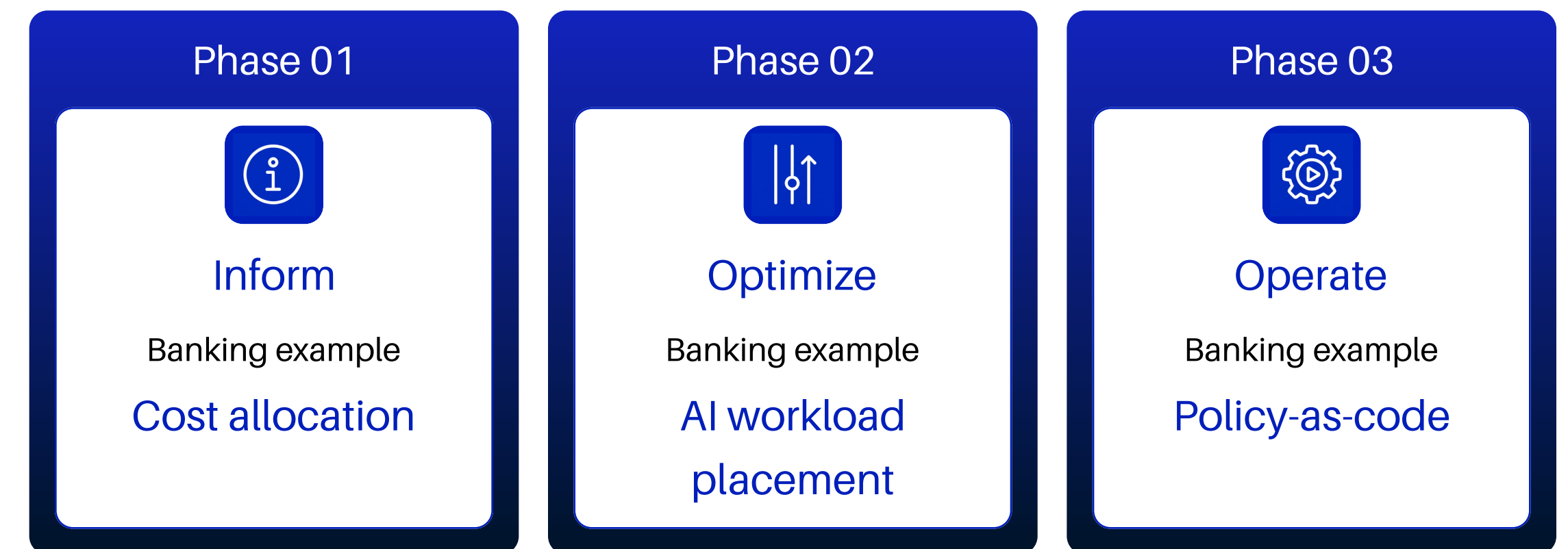
What Is FinOps Maturity?

FinOps maturity is the ability to use cost, usage, ownership and outcome data to make better technology decisions. In banking, mature FinOps connects Cloud+ visibility with governance, automation and business value, so teams can manage spend across cloud, SaaS, private infrastructure, data center and AI workloads.

The FinOps lifecycle still matters: Inform, Optimize, Operate. Banks first need to understand what's being spent, where it's happening and who owns it. They can then identify where spend should be reduced, redirected or better matched to demand. Finally, those decisions need to become everyday practice through policies, controls, automation and team enablement.

In reality, those phases don't happen neatly one after another. They run at the same time. One AI workload might still need basic visibility, while a SaaS estate is ready for optimization and a cloud landing zone already needs automated guardrails.

FinOps Phases in Banking



The direction is clear. [The State of FinOps 2026 report](#) says 98% of FinOps teams now manage AI spend, up from 31% two years earlier. AI is no longer an edge case for FinOps. It's becoming part of the core discipline.

Introducing The Cross-Functional FinOps Operating Model For BFSI Organizations

In a bank, a single technology spend decision can cross several functions before anyone signs it off. The line of business wants the capability. Engineering builds it. Finance owns the budget. Product owns the business case. Procurement negotiates the contract. Risk and Compliance review the exposure. Security checks data handling. Model Risk Management reviews the model.

When AI is involved, the handoffs multiply. Usage-based costs mean the budget is moving in real time, while model behavior, customer data, vendor concentration and data residency all need to be understood before the workload reaches production.

No single department can own that decision alone. Cross-functional governance is the only model that works, and the FinOps Practitioner is the connective role.

The Roles That Need To Be At The Table

A banking FinOps operating model should include:



FinOps Practitioner:

owns the day-to-day operating model, reporting logic, allocation rules, optimization roadmap and governance cadence. This role bridges Engineering, Finance and Product



Leadership:

sets strategy, budget envelopes and accountability. In banking, this usually means CTO, CIO, COO and CFO involvement, with one named executive sponsor



Engineering:

builds and runs the services. Cost becomes part of good engineering practice, not an afterthought

**Finance:**

owns budgeting, forecasting, showback, chargeback and financial reporting

**Product:**

connects spend to customer, channel, product and P&L outcomes

Banks also need the control functions in the room:

**Risk and Compliance:**

vendor concentration, operational resilience and regulatory evidence

**Model Risk Management:**

validation, monitoring and approval of models that affect customer, fraud, credit or operational outcomes

**Security:**

data residency, encryption, access and segregation

**Procurement:**

enterprise commitments, vendor terms and contract risk

A Worked Example: AI-Driven Fraud Detection

Picture a retail banking team that wants to scale an AI-driven fraud-detection pipeline.

The business case is strong. The team expects lower fraud losses, fewer false positives and a better customer experience. Engineering scopes the extra compute, storage, inference and API usage, while the FinOps Practitioner stress-tests the forecast against similar workloads. They compare the expected cost per transaction with the current fraud process and show how different vendor, region or commitment choices would change the economics.

Finance then models the net P&L impact. Model Risk Management reviews the model approach, monitoring and challenger-model arrangements. Security checks whether customer data can be processed in the proposed environment and region, while Procurement reviews vendor terms, discounts and concentration exposure.

Leadership can then approve the investment with review cycles and escalation thresholds already in place.

The FinOps Practitioner is what stops this becoming a chain of disconnected handoffs. Without that role, the request either slows down at every checkpoint or moves ahead without one team seeing the full cost, risk and value picture.

Simplified RACI matrix — FinOps operating model

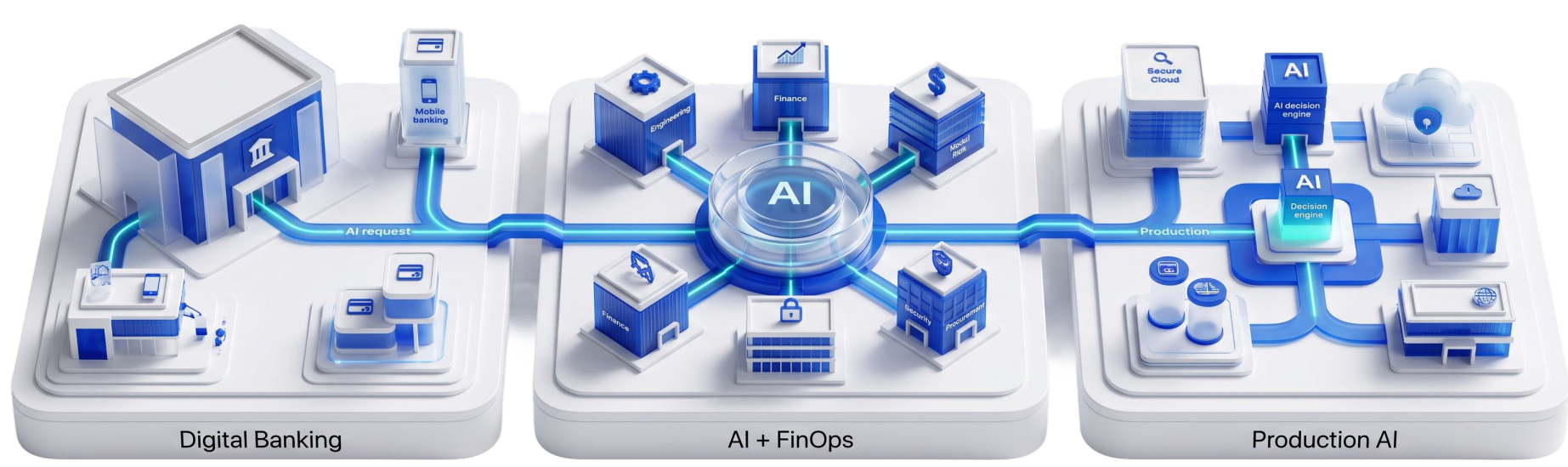
	 Leadership	 FinOps practitioner	 Engineering	 Finance	 Product	 Risk & compliance	 MRM
Strategy	A	R	C	C	C	C	I
Forecasting	I	R	C	A	C	I	I
New AI spend approval	A	R	C	R	C	C	C
Cost optimization	I	A	R	C	C	I	I
Governance and policy	C	R	C	I	I	A	C
Audit trail	I	R	C	I	I	A	C

 Accountable

 Responsible

 Consulted

 Informed



Where The CCoE & AI CoE Fit

The FinOps practice shouldn’t run in isolation. It usually sits inside, or works closely with, a Cloud Center of Excellence. The CCoE provides the standards, reference architectures and engineering support that make Cloud+ FinOps practical at scale.

For AI workloads, the same logic applies to the AI Center of Excellence. The AI CoE owns model lifecycle standards, responsible AI policies, approved model lists, evaluation thresholds and human-in-the-loop rules.

The FinOps Practitioner draws on both. That’s how a single AI workload can be reviewed through cloud-platform, model-lifecycle, cost, value and risk lenses before it goes live.

Which Technical Features Define The Cloud+ FinOps Operating Model?

A Cloud+ FinOps practice needs more than a dashboard. It needs enough structure to make spend visible, comparable and enforceable across the full banking estate.

The technical stack doesn't need to be overcomplicated. Its job is to answer four questions:



What are we
spending?



Who owns
it?



What value is
it creating?



Which controls
apply?

Data Ingestion & Normalization

The foundation is timely cost and usage data from every place spend occurs. For a large bank, that could include hyperscaler billing, SaaS invoices, private cloud chargeback, data center cost models, AI vendor billing, model gateways and ERP data.

Monthly totals aren't enough. Banks need cost data that can be traced to products, workloads, teams, environments, models and regions.

Open standards such as FOCUS can help by giving cost and usage data a more consistent format across providers. FOCUS 1.3 added support for contract commitments, split cost allocation and data-quality dimensions such as recency and completeness.

That matters because Cloud+ FinOps breaks down quickly when every vendor reports cost in a different way.

FinOps Platforms & ERP Integration

Most large banks will use a mix of native cloud tools and third-party FinOps platforms. Native tools provide depth inside one provider. A FinOps platform creates the wider Cloud+ view across cloud, SaaS, AI and private infrastructure.

That view also needs to tie back to the financial system of record. If allocated costs in the FinOps platform don't reconcile with the ERP, the audit trail weakens. Finance may see one number, Engineering another, and Risk a third.

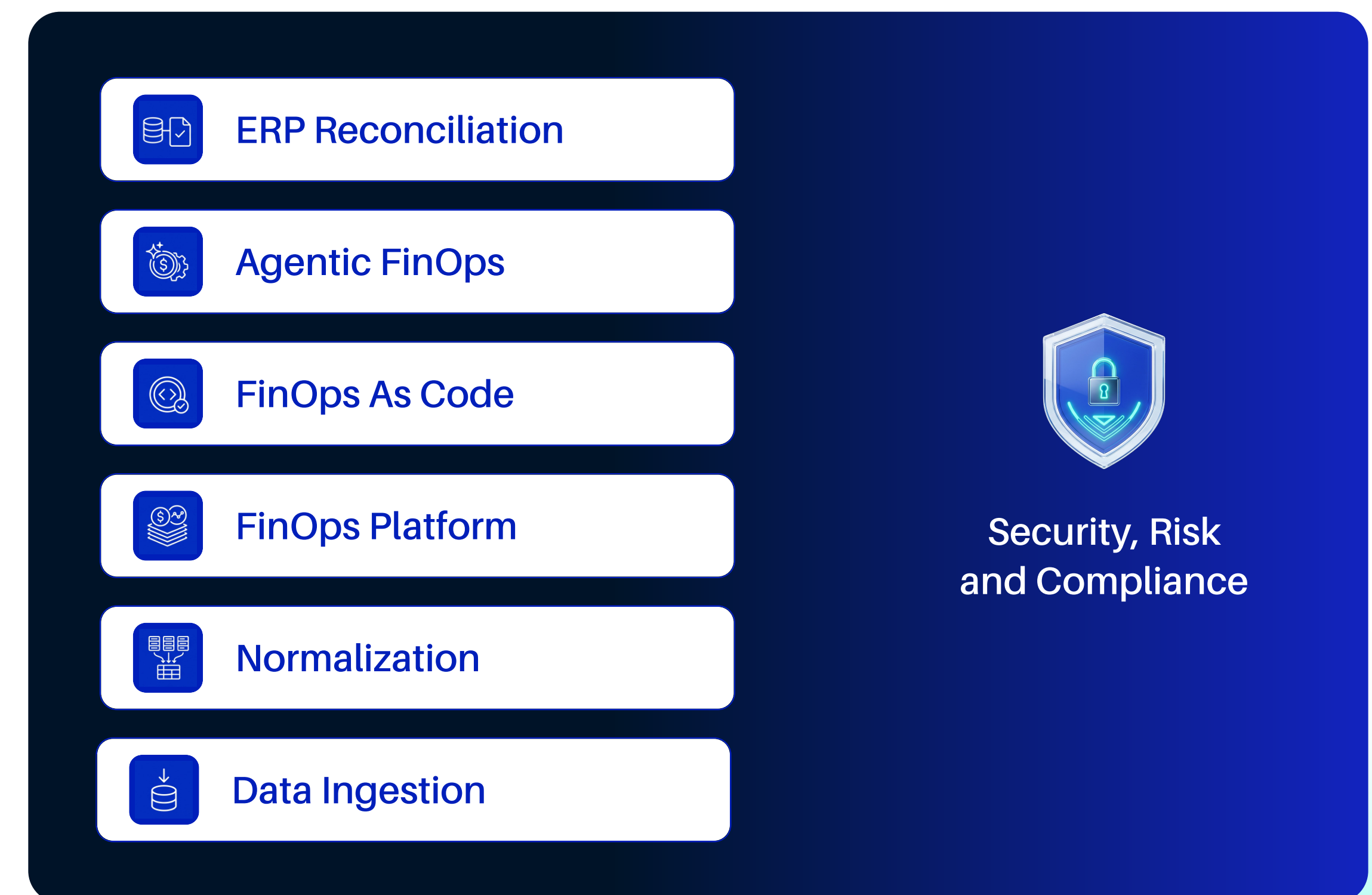
In a bank, that's not sustainable.

FinOps As Code

FinOps as Code embeds financial and governance policies into the engineering workflow.

Untagged resources can be blocked before deployment. Non-production environments can be shut down outside agreed hours. EU customer workloads can be prevented from landing in non-approved regions. AI usage can be controlled through approved model lists, routing rules and budget thresholds.

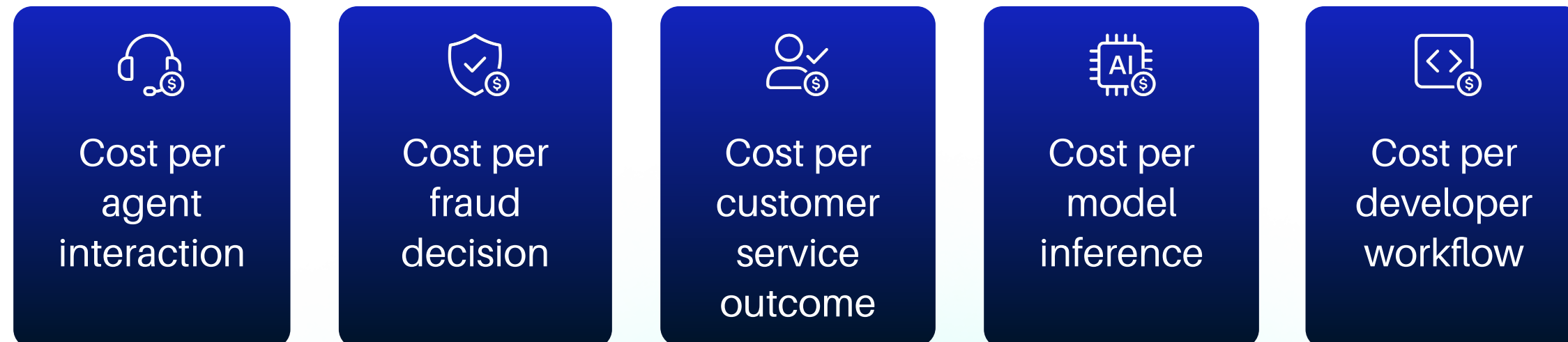
This is where policy stops being a document and becomes something teams work within every day. It also changes the role of FinOps. The function isn't only reporting what happened last month. It's shaping what teams can deploy tomorrow.



How Agentification & Agentic FinOps Work In BFSI

Agentification is already changing banking workflows. Customer service bots, internal workflow agents, collections assistants, fraud tools and developer co-pilots all create dynamic usage, with each interaction, prompt, model route and API call carrying a potential cost.

FinOps helps by giving banks the unit economics behind those workflows:



But AI agents can also support FinOps directly. An agent can monitor spend, spot anomalies, flag policy breaches, recommend fixes and, in some cases, carry out low-risk actions automatically.

That's useful, but it's also sensitive in a regulated bank. A low-risk action might be shutting down idle non-production resources outside business hours, applying a missing tag or raising an alert when AI inference spend moves outside an agreed threshold.

It shouldn't mean resizing a production database without approval. It shouldn't mean changing commitment levels automatically. It shouldn't mean changing how a production model behaves without Model Risk Management sign-off.

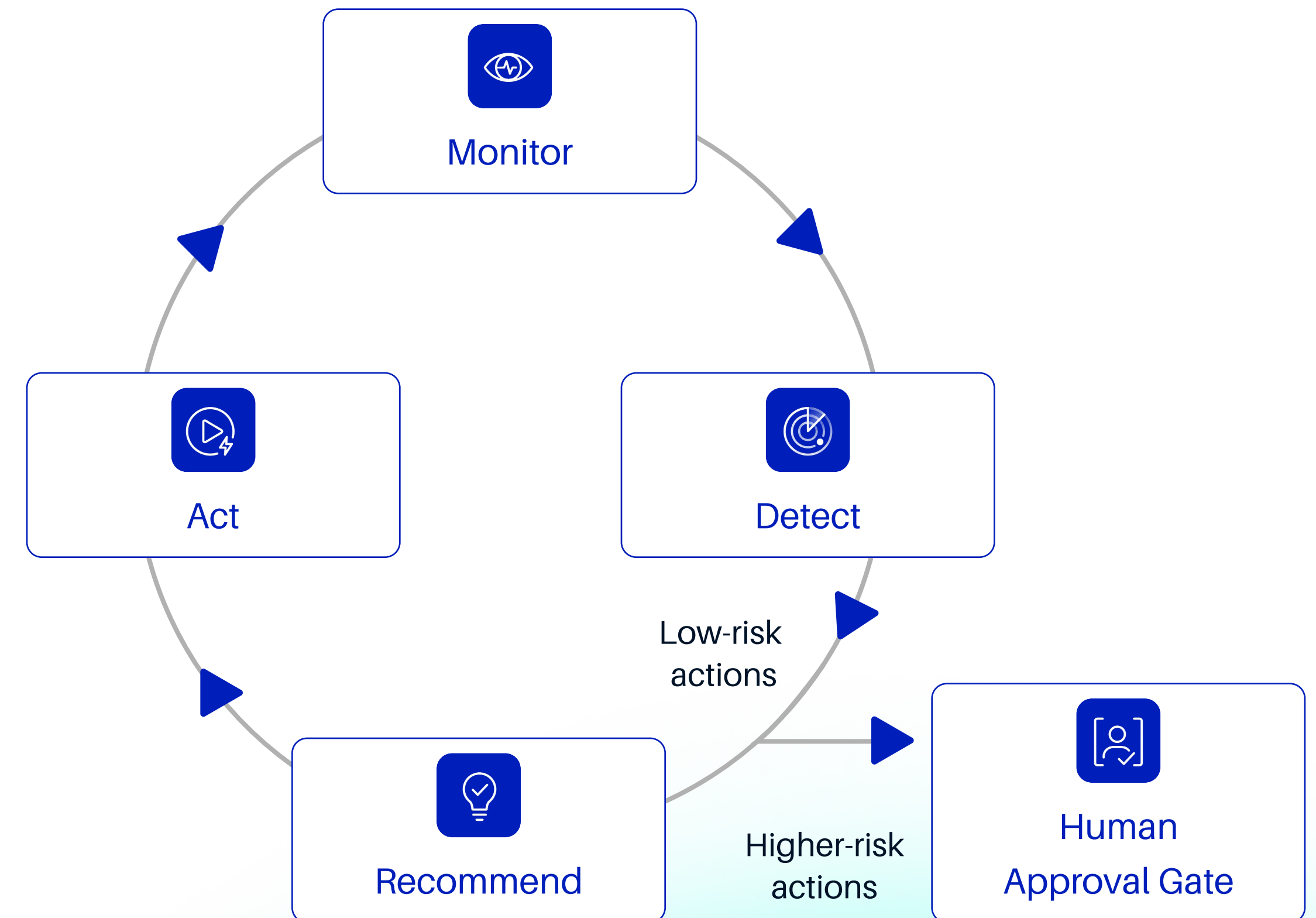
How Agentification & Agentic FinOps Work In BFSI

The boundaries need to be clear:

-  Agents act only below defined value and risk thresholds
-  Every observation, recommendation and action is logged
-  Recommendation, approval and execution remain separate
-  Every automated action has a tested rollback route

This is where FinOps maturity becomes visible. Less mature teams use AI to explain last month's spend. Mature teams use governed automation to stop the wrong spend from happening in the first place.

AI-Driven Workflow Control



How AI Is Applied To Banking FinOps: Real-World Signals

The public banking examples don't usually describe themselves as FinOps deployments. That's fine. Their value is different. They show banks scaling AI and measuring outcomes, which is the test any Cloud+ FinOps practice eventually has to pass.

BBVA

Enterprise AI Adoption At Scale

BBVA is expanding ChatGPT Enterprise

to 120,000 employees across 25 countries

through a multi-year AI transformation program with OpenAI. From a FinOps perspective, this turns AI from a software rollout into an operating model question: who owns usage, how value is measured, and which use cases should scale.

citibank

Agentic AI Inside A Banking Platform

Citi has introduced

Agentic AI into Citi Stylus Workspaces.

its proprietary AI platform. For FinOps teams, agentic AI changes the unit economics. The question is no longer only what a model costs to run. It's what an agentic workflow costs per completed task, decision or customer outcome.

The common thread is measurement. Banks need to know which AI use cases improve productivity, reduce fraud, increase resilience or improve customer experience. The rest only add cost.

In some cases, the value created can help fund the next wave of transformation. That's the position every CFO, CTO and FinOps Practitioner is trying to reach: AI investment governed closely enough that value can be measured, defended and reinvested.

Regulation, Risk And Resilience: FinOps Under The Supervisors' Gaze

Regulators don't mandate FinOps. They do expect banks to understand their technology estate, control third-party relationships, manage operational resilience and produce evidence when asked.

[DORA](#) has applied across the EU since January 17, 2025, strengthening digital operational resilience requirements for financial entities and ICT third-party service providers. In the UK, [PRA SS2/21](#) sets expectations for outsourcing and third-party risk management across PRA-regulated firms.

FinOps supports that environment because it creates an evidence layer behind technology spend. A mature Cloud+ FinOps practice can give risk and compliance teams six practical capabilities:



A unified vendor register:

active technology vendors, committed spend, regions, workloads and ownership



Concentration-risk visibility:

dependency by provider, region, service category and AI vendor



Workload-placement governance:

evidence of where workloads run, why they run there and what it would take to move them



Row-level audit trail:

a traceable route from source bill to allocation rule, approval, policy and reporting output



Cost-to-outcome attribution:

cost per product, transaction, inference, model, journey or business outcome



Policy-as-code enforcement:

controls embedded into delivery workflows, not left in static policy documents

Regulation, Risk And Resilience: FinOps Under The Supervisors' Gaze

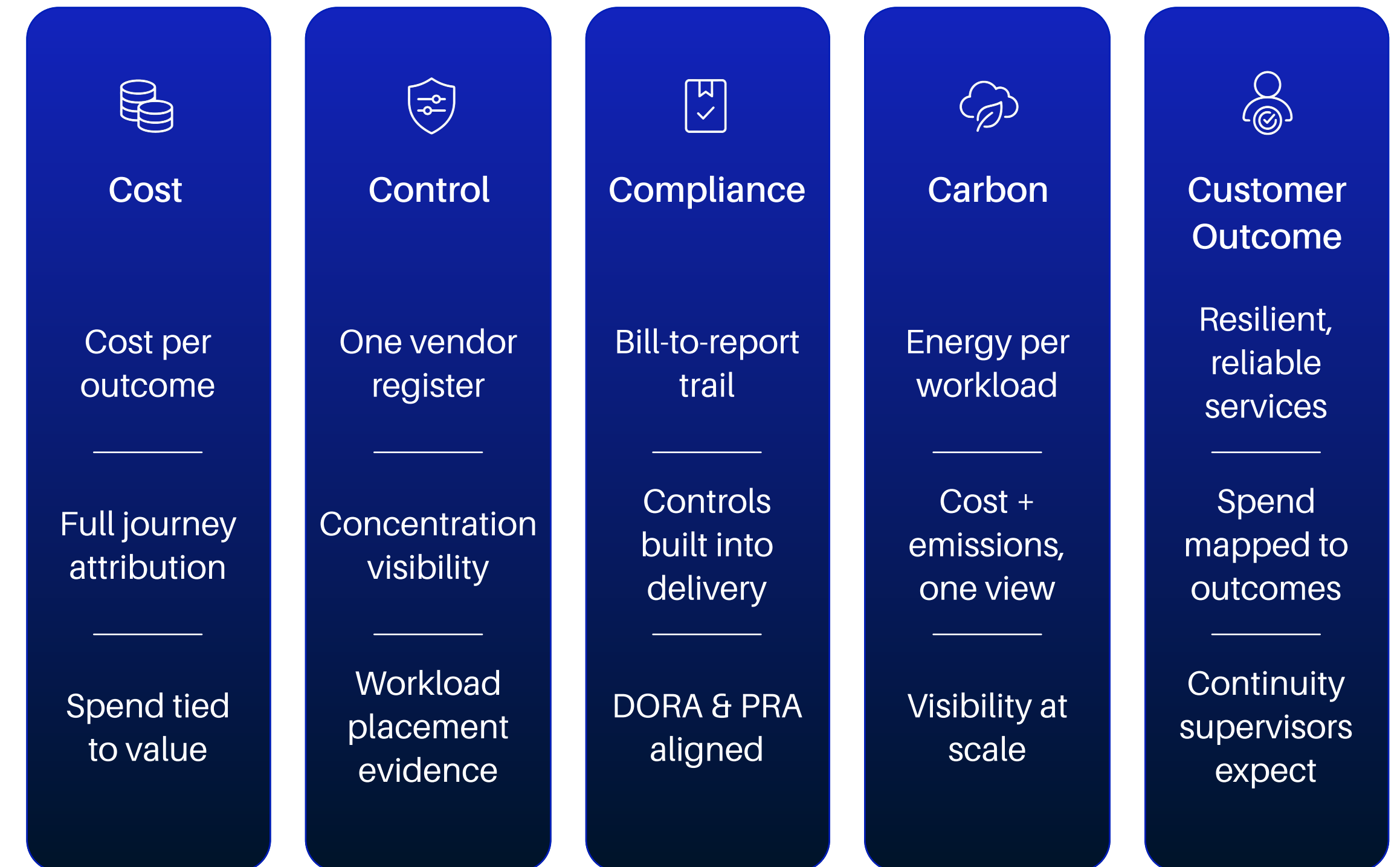
This is the practical compliance value of FinOps. It doesn't replace Risk, Security, Procurement or Model Risk Management. It gives those teams stronger evidence and better controls.

It also supports sustainability. As AI workloads grow, banks will need clearer visibility into the cost and energy profile of different workload-placement decisions. FinOps gives teams a cleaner way to connect efficiency, cost and emissions data at workload level.

With AI moving quickly, regulatory expectations will continue to develop. But the core requirement is already clear: banks need to show that AI and technology spend is controlled, explainable and tied to resilient services.

Pillars of Responsible AI

How We Integrate Ethical Practices Into AI Spending



Your Next Steps: Introducing Ciklum PRODIGY

The key to unlocking the potential of FinOps in BFSI is coordination: having a clear strategy for cloud, automation and AI in place, and an end-to-end strategy for building, deploying and scaling AI in the right ways.

That's where Ciklum [PRODIGY](#) comes in.

PRODIGY is Ciklum's unified platform and methodology, designed to bring control, order and measurable business outcomes into AI deployments. It involves a range of vertical, horizontal and technical accelerators that we've developed to support faster implementation, greater personalization and stronger differentiation in a competitive market.

In the context of FinOps, this includes FinOps as Code and agentic FinOps that enable monitoring and optimization under predefined guardrails. These agents, tailored to your specific needs through a combination of our expertise and proprietary technology, can help you meet the twin goals of cost optimization and clear compliance.

The first step on this journey is to take part in our free, no-obligation FinOps assessment, which will provide clarity on your current position and help you understand the best FinOps implementation steps for you.

Contact our team today to

[book your assessment](#)

or to find out more on

[Ciklum PRODIGY.](#)

Our Experts



Aaron Jones

Sector Lead, BFSI

✉ aaron.jones@ciklum.com



Maksym Lypivskyi

AI R&D Director

✉ maksym.lypivskyi@ciklum.com