

Artificial Intelligence

## From Rate Cards to Outcomes: Consulting's Fourth Transformation

Raja Pabba

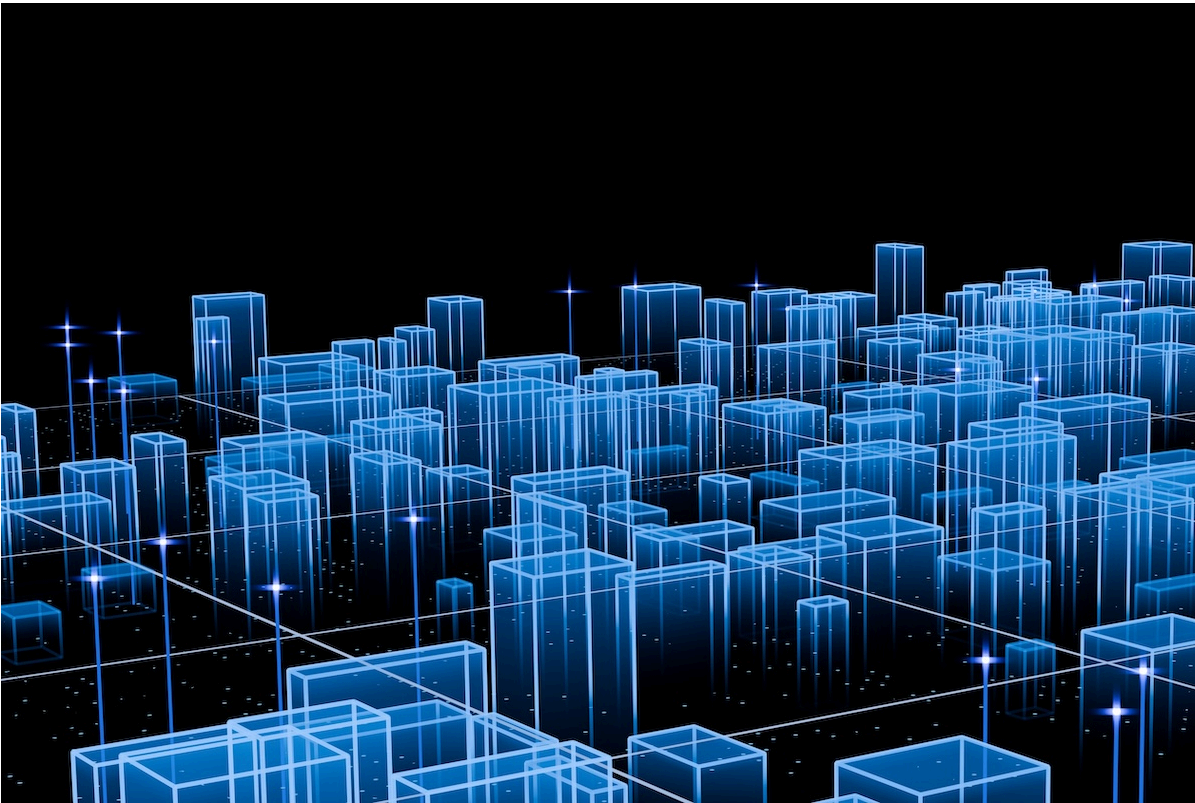


Image Credit | PW.Stocker

*Why AI demands a new model – and why 'Service as a Software' will define the winners*

**The Argument:** Professional services firms have survived three technology-driven transformations: ERP implementation (1990s), web/mobile enablement (2000s), and SaaS/cloud platforms (2010s). Each wave changed what clients bought, how they paid, and what they received – but left the fundamental consulting model intact. Clients still purchased human expertise, measured in hours or FTEs.

### **AI breaks this pattern.**

The fourth transformation inverts the services model entirely. Rather than software enabling consultants to work faster, expertise itself becomes software. I call this **‘Service as a Software’** – the encoding of domain judgment into autonomous systems that deliver outcomes directly, with humans supervising rather than executing.

This article argues that the critical capability for this era is not AI engineering but Expertise Architecture: the systematic methodology for capturing domain judgment and encoding it into machine-executable reasoning. Firms that master this capability will capture disproportionate value. Those that treat AI as merely another accelerator for existing labor models will find themselves disrupted by focused entrants who start without legacy economics to protect.

## **The Four-Era Framework**

Each technology wave transformed consulting economics in predictable ways:

<b>Era</b>	<b>Tech Driver</b>	<b>What Clients Buy</b>	<b>How They Pay</b>	<b>What They Get</b>
<b>1</b>	<b>ERP</b>	<b>Staff augmentation</b>	<b>Rate card</b>	<b>Capacity</b>
<b>2</b>	<b>Web / Mobile</b>	<b>Tech-enabled services</b>	<b>FTE-based</b>	<b>Skills + accelerators</b>
<b>3</b>	<b>SaaS / Cloud</b>	<b>Asset-based platforms</b>	<b>FTE + transaction</b>	<b>Tools + frameworks</b>
<b>4</b>	<b>AI / GenAI</b>	<b>Service as a Software</b>	<b>Outcome-driven</b>	<b>Machine-led outcomes</b>

# The Incumbent Response: Two Paths, Same Gap

The industry's largest players recognize the shift — but their responses reveal an unresolved strategic tension.

On January 22, 2026, McKinsey and AWS launched the Amazon McKinsey Group (AMG), a joint venture explicitly designed around outcome-based pricing. The structure is notable: rather than billing for consultant hours, AMG ties fees to measurable transformation results on engagements exceeding \$1 billion. This is a structural bet that the traditional labor model cannot survive Era 4. McKinsey is not adding AI to consulting; it is repositioning consulting around AI-enabled delivery.

Yet even this bold move exposes the gap. AMG still depends on McKinsey consultants to interpret client context, design transformation roadmaps, and validate AI-generated recommendations. The 'expertise layer' remains human. The joint venture changes who bears performance risk, but it does not fundamentally change how expertise gets delivered. McKinsey has restructured the economics without yet encoding the judgment.

Contrast this with Accenture's approach. The firm announced \$3 billion in AI investments and has built impressive technical capabilities — AI factories, proprietary tools, thousands of trained practitioners. But the underlying delivery model remains intact: consultants use AI to work faster, clients still pay for FTEs, and value is measured in hours saved rather than outcomes achieved. This is Era 3 optimization, not Era 4 transformation. AI augments the labor model; it does not invert it.

Deloitte, EY, and others occupy similar positions — significant AI investment, genuine technical capability, but strategic ambiguity about whether AI is a tool for consultants or a replacement for consulting. The ambiguity is rational: these firms generate billions in labor revenue. Protecting that revenue while simultaneously enabling autonomous delivery creates organizational tension that no amount of investment resolves.

**The pattern across incumbents is consistent: recognition without resolution.** They see the shift. They are investing heavily. But none has cracked the core problem — systematically encoding the domain expertise that makes their consultants valuable. Until they do, they remain vulnerable to focused entrants who start without legacy economics to protect.

## Where Service as a Software Already Works

While incumbents navigate strategic tension, a new category of company demonstrates that ‘Service as a Software’ is not theoretical — it is already operating.

Vertical AI companies target functional domains where decision patterns are bounded, judgment is repeatable, and outcomes are measurable. Rather than building general-purpose AI tools, they encode domain-specific expertise into autonomous systems that deliver results directly. The model works because these companies start with encoded expertise as their core asset, not as an enhancement to labor.

Consider the pattern: A procurement AI platform doesn’t just surface contract anomalies for humans to review. It encodes the judgment that procurement professionals apply — what constitutes a meaningful price variance, which suppliers warrant scrutiny based on risk profile, when to escalate versus auto-approve. The system doesn’t accelerate human work; it executes the work with human oversight.

In my own experience building an AI-native FinOps platform, I’ve observed how this plays out operationally. FinOps — the practice of managing cloud and technology spend — involves hundreds of decisions daily: Is this cost spike an anomaly or expected? Should this workload be rightsized? Does this variance warrant executive attention?

Traditional approaches surface data and expect humans to decide. Our approach encodes the decision logic itself. We built what we call PRISM — a methodology that decomposes FinOps into five decision domains (Proactive monitoring, Resource optimization, Infrastructure management, Spend economics, and Management governance), each with explicit thresholds, escalation rules, and context-dependent reasoning. The AI doesn’t just

detect a 15% spend increase; it knows that 15% in production environments during quarter-end is expected, while 15% in development environments on weekends warrants immediate investigation.

**The result: decisions that previously required analyst review now execute autonomously within defined risk parameters.** Humans supervise exceptions rather than processing routine judgments. Time-to-action compresses from days to minutes. And critically, the value delivered is measurable in outcomes — cost avoided, efficiency gained — not hours worked.

This pattern — bounded domain, encoded judgment, autonomous execution with human oversight — defines where ‘Service as a Software’ already works. The question for incumbents is whether they can replicate it before these focused players expand.

## Where It Fails: The Encoding Gap

For every successful vertical AI deployment, there are dozens that stall — not because the AI doesn’t work, but because the expertise was never encoded.

Early in our product development, we learned this lesson directly. We deployed anomaly detection to flag unexpected cost variances in cloud infrastructure. The model performed well technically — it identified patterns humans missed, processed data at scale, and surfaced potential issues in real time. By any AI benchmark, it succeeded.

### **But in practice, it failed.**

The system flagged everything that deviated from baseline: a 12% increase in compute spend, a new storage allocation, a spike in data transfer costs. Within the first week, it generated over 200 alerts. Finance teams, already stretched thin, couldn’t process them. They had no way to distinguish signal from noise because the system didn’t encode what practitioners know — that a 12% increase during product launch is expected, that new storage allocations tied to approved projects aren’t anomalies, that data transfer spikes during backup windows are routine.

Without encoded thresholds, context rules, and escalation logic, the AI created more work, not less. Alert fatigue set in within two weeks. Teams began ignoring notifications. By month two, the anomaly detection was effectively shelfware — technically operational, practically abandoned. We had automated data processing but not decision-making.

This pattern repeats across enterprises. A legal team deploys AI to review contracts; without encoded risk tiers, lawyers still review 100% of outputs. A customer service team launches an AI assistant; without encoded resolution paths, 60% of queries escalate to humans. A finance team automates expense auditing; without context rules, 80% of flags are false positives.

**The failure mode is consistent:** organizations deploy AI capability without encoding the judgment that makes capability useful. They automate the process layer — data ingestion, pattern recognition, alert generation — while leaving the expertise layer untouched. The result is augmentation at best, shelfware at worst.

## Why It Fails: The Missing Layer

The failures above share a common root cause. Enterprises have invested decades in process frameworks — APQC taxonomies, BPMN models, SIPOC documentation. These capture workflow: what activities happen, in what sequence, with which roles. They do not capture judgment.

Consider a common process step: ‘Review budget variance.’ A process model shows this as an activity box connected to a role. But what AI needs to automate this step is entirely different: When is a variance significant? 5%? 10%? Does it depend on the cost center? The time of year? The trend direction? These are judgment calls that exist in practitioners’ heads but not in any process documentation.

**This is the critical gap: AI can automate process. AI cannot automate judgment without encoding.**

I call this missing layer Expertise Architecture — the systematic encoding of domain judgment into machine-executable form:

Layer	What It Captures	Status
<b>Expertise Architecture</b>	Judgment, thresholds, heuristics, context-dependent reasoning	<b>MISSING — AI-ready layer</b>
Process Model	Activities, sequence, roles, handoffs (BPMN, SIPOC)	Commoditized
Process Classification	Taxonomy, standardized naming (APQC)	Commoditized

**The strategic implication is clear:** Process classification and process models are commoditized — everyone has them. The scarce, ownable capability is methodology for building the Expertise Architecture layer. Firms that build this layer first will define the category.

## What To Do: Guidance for Three Audiences

**For consulting firm leaders:** Current IP libraries are necessary but insufficient for Era 4. The strategic question shifts from ‘how do we deploy AI tools’ to ‘how do we encode our expertise before others do.’ Protecting existing labor revenue will delay transformation; focused entrants face no such constraint. The McKinsey-AWS model shows one path — restructure economics around outcomes. But without encoded expertise, outcome-based pricing shifts risk without changing capability.

**For enterprise buyers:** Evaluate vendors on expertise encoding methodology, not AI capability alone. Ask: ‘How do you capture and validate domain expertise in your AI systems?’ Demand transparency on human-AI task allocation. If a vendor’s AI still requires your team to review every output, you’re buying augmentation, not transformation. Expect and negotiate for outcome-based pricing as the standard — and verify the vendor has encoded enough judgment to deliver on it.

**For new entrants:** Functional verticals (FinOps, procurement, revenue operations) offer clearer encoding targets than industry verticals. The bounded nature of these domains — repeatable decisions, measurable outcomes, defined thresholds — makes expertise

encoding tractable. First-mover advantage accrues to firms that establish trust and governance frameworks. The window for category definition is open but will close within 3-5 years as incumbents resolve their strategic ambiguity.

## Conclusion

The consulting industry has survived three technology transformations by adapting its delivery model while preserving its fundamental economics: clients pay for human expertise. AI breaks this pattern because it enables expertise itself to become software.

The winners of this transformation will not be determined by AI capability – that is rapidly commoditizing. They will be determined by who solves the expertise encoding problem first. The firms that build Expertise Architecture – the systematic methodology for converting domain judgment into machine-executable reasoning – will capture disproportionate value. Those that treat AI as another tool for consultants will find themselves disrupted by focused players who started without legacy economics to protect.

**The shift from rate cards to outcomes is not a pricing change. It is an architectural change.** And the architecture that matters most is not technical – it is the encoding of human expertise into systems that can act on it.



Raja Pabba [Follow](#)

Raja Pabba is Founder and CEO of CloudMetrics, building an AI-native FinOps platform. He spent 25 years at EY and Accenture advising Fortune 500 clients on technology strategy. He holds an MBA from the University of Chicago Booth School of Business and has a provisional patent on expertise encoding methodology.