Guide to Operationalizing FinOps

How to better optimize cloud costs in the age of AI

IBM

Contents

01

What is FinOps?

Today, organizations are challenged to fund innovation while responsibly managing cloud costs and prioritizing spend for maximum business value. When success is measured by uptime and high customer satisfaction scores, generative AI (gen AI) can dramatically change the game. Yet predicting the true cost of gen AI computing power is difficult—especially when IT executives find themselves having to boost their cost estimates for gen AI by more than 3 times in just 4 months.[1]

Yet even though budgets are tight, many IT teams err on the side of caution when it comes to managing their applications, with or without gen AI. To mitigate performance risks—which can negatively affect the customer experience and those critical satisfaction scores—developers often overprovision resources. This common practice helps explain why IT executives estimated that 24% of their cloud spend was wasted in 2023.[2]

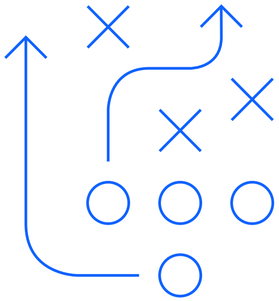
That's where FinOps can help. Financial operations (FinOps) is the cloud financial management practice designed to bring together engineers, financial teams and business leaders to give organizations greater insights into their technology spend, eliminating waste in the process.

More precisely, the FinOps Foundation, an organization founded to advance the people who manage the value of cloud, puts it this way: "FinOps is an operational framework and cultural practice which maximizes the business value of cloud, enables timely data-driven decision-making and creates financial accountability through collaboration between engineering, finance and business teams."[3]

In this ebook, we'll discuss the basics of a FinOps program and the need for intelligent automation. We'll look at how gen AI can be added to a FinOps practice for even greater value and how to overcome typical FinOps blockers.

02

Begin with the basics



The FinOps framework is robust with capabilities that help control spend, such as allocating costs or optimizing workloads. However, if you're early in your practice or uncertain of where to begin, here are 3 basic steps to better understand and manage your cloud spend.

1. Start with cost visibility and hold IT teams accountable

Too often, IT teams expect failures—an expectation that forces them to over allocate resources as a means of mitigating risk. Or cloud engineering and ITOps teams monitor their environment until they receive an alert that they've gone over budget or that performance has degraded. However, neither expected failures nor a reactive approach is an effective use of time and resources, both of which are too often in short supply.

Instead, use the FinOps framework to gain spending visibility of the entire cloud environment. Work across the organization to understand and document allocation of cost centers and teams, budgeting and forecasting, and chargebacks and showbacks. That visibility gives you the necessary insights to proactively know where spend is going and where it can be reallocated for optimal results.

2. Make cost optimization your next priority
Once you've gained a greater understanding of where money is going, you can begin to make adjustments. Cost optimization in the context of FinOps is the exercise of discovering and acting on opportunities for cost savings—either by reducing the amount you use or the price that you pay.

– **Optimizing what you use**
Engineers can optimize and reduce cloud compute, storage, database as a service (DBaaS) and Kubernetes through automated actions, such as rightsizing instances, deleting or suspending idle instances, or dynamically scaling workloads or placements. Vertical and horizontal scaling actions are based on real-time application demand, while automation helps ensure that virtual machines, containers, databases and storage are appropriately sized and horizontally scalable applications are elastically scaled to meet demand.

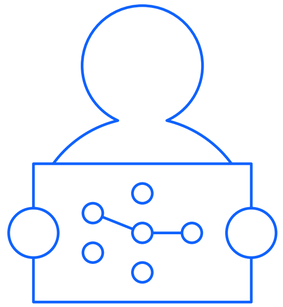
– **Optimizing what you pay**
Cloud engineering and ITOps teams can reduce the price they pay using commitment-based discounting. This discounting approach includes reserved instance scaling actions to increase inventory utilization or purchase more reserved instances, which allows you to maximize reservation-to-virtual-machine coverage. In addition to optimizing what you pay, you can also reduce manual analysis, giving valuable time back to your teams.

Be prepared, however, for some hesitancy from your teams. Cloud engineers need to trust that neither performance nor end-user experiences will be in jeopardy if they take their hands off the "virtual wheel." Your teams won't take cost optimization actions—let alone automate them—if there's even a hint of risk to performance. The only way to confidently implement automation is through software that takes a performance-first approach, accounting for the entire application stack and all the resource dependencies across the infrastructure it runs on.

3. Build automation into your cost optimization programs
Manual execution can reduce cost and improve efficiency in isolated exercises. But continuous cost optimization at scale is only possible with automation. Automation is key to effectively scaling and managing the hundreds, if not thousands, of cost optimization actions that need to be taken in real time. It's also the best way to take advantage of cloud elasticity, which is the ability to dynamically adjust resource configurations, so applications get exactly the resources they need, when they need them. Then they can be readjusted when those resources are no longer necessary.

03

Infuse gen AI at every phase



Once you have a deeper visibility into spend and a solid optimization plan, consider how you can use FinOps beyond basic cost management and bring gen AI into the process. Common applications of gen AI with FinOps include automated reporting, predictive analytics, compliance monitoring and cost optimization, resulting in more value with less effort.

More specifically, gen AI can be applied at each of the 3 phases that FinOps practitioners typically work through: inform, optimize and operate.

Inform: Visibility and allocation
In the inform phase, all stakeholders are empowered with the information and understanding they need to make informed, cost-effective decisions around cloud usage, including This includes cost allocation discovery and cloud program total cost of ownership (TCO).

By applying gen AI in the inform phase, users can gain an understanding of how applications are consuming resources and services across an organization's cloud environment for better forecasting and benchmarking. Using a conversational experience and natural language processing (NLP), you can ask about spend trajectories and anomalies without having to dig through dashboards.

Optimize: Rates and usage.
In the optimize phase, practitioners identify opportunities for additional savings and improved cloud efficiency. This process includes right-sizing recommendations and preferences, maximizing reserved instances and savings plans, and automating financial commitment instruments.

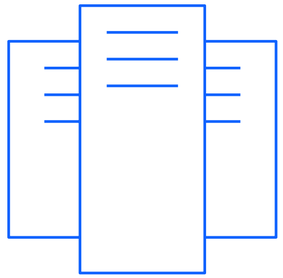
With gen AI in the optimize phase, cloud users can take what they've learned and apply it for additional savings. Once again, using NLP, gen AI can help provide insights into spend and detect patterns and trends.For example, the user could ask, "Why is spend out of line with expectations?" Gen AI can also provide a high level of granularity in meeting budget key performance indicators (KPIs), improving accuracy and reliability.

Operate: Continuous improvement and usage.
In the operate phase, practitioners evaluate performance against business objectives. Then they look for ways to implement organizational changes to improve and operationalize their FinOps practices. This process includes basic budgets and forecasting, workload planning and reporting.

With gen AI in the operate phase, users can automate their FinOps practices and build a culture of continuous improvement. Gen AI can detect patterns at a much more refined level. You can ask, for example, about seasonality of spend, the costs of software licenses or labor, and other more granular aspects of costs. You'll then be equipped to show exactly what you're spending and why, and tie budgets directly to business outcomes.

04

Less waste, more sustainable IT

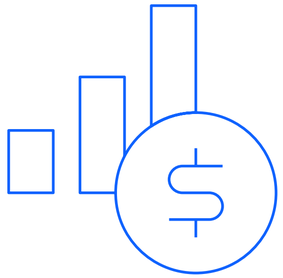


In 2022, data centers accounted for 2% of all global electricity usage. The rapid growth of gen AI—and the corresponding computing power and cooling required—means that greater demand is expected through 2026. As you build your FinOps practice, it's important to recognize the impact you can have on more than just the bottom line. Of course, responsibly managing resources is top of mind. IT budgets—already notoriously tight—are under even more pressure with gen AI computing added in. But now sustainability is also high on the agenda. In one recent survey, 65% of C-level executives interviewed said it's one of the top 3 priorities today, compared to just 28% 3 years ago. This same need for better resource management is born out in the 2024 State of FinOps report, too, where reducing waste is the #1 priority.[3]

Of course, reducing waste and creating more sustainable cloud management systems require a commitment that extends well beyond a FinOps practice. They require a holistic approach that accounts for emissions and pollutants emitted through various operations specific to different businesses. However, the goals of FinOps often overlap with those of sustainability programs. With a focus on making the most of every dollar spent, FinOps can help drive the adoption of automated dynamic resourcing. This focus makes it possible to realize maximum efficiency in hybrid and multicloud estates in an immediate and impactful way, because responsibly operating in the cloud materially reduces your carbon footprint—a win-win opportunity to reduce both cost and emissions.

05

Longer term: Create a FinOps for all approach



For most FinOps practitioners, the goal is often to increase the value of cloud investments while ensuring high performance of applications. In short, it's to make cloud a competitive advantage for your organization. Longer-term goals should enable you to:

- Gain transparency into all cloud spend and take control of cost overages.
- Manage workloads and costs consistently across multiple clouds.
- Operate at the lowest cost possible without having to worry about end-user experiences.
- Deploy automation that dynamically scales and resizes workloads to optimize resource consumption.
- Reduce costs immediately and continuously by ensuring workloads only consume what they need to perform.
- Automate dynamic resourcing actions so applications, and the infrastructure they run on, continuously manage to service-level objectives (SLOs) that correlate to business success.
- Optimize cloud migrations through application-driven planning capabilities, asking which workloads are best suited for the cloud for efficient consumption from the start.

As your FinOps practice matures and you adopt more cloud-native applications, your approach can become more holistic. A FinOps for all approach delivers capabilities for all personas—IT, finance and engineering—to optimize cloud resources for speed, cost and quality. This approach includes the following:

- Shift left capabilities to help engineers proactively optimize and automate actions that prevent cost issues before they arise
- Tailored views and flexible reporting to keep leadership teams informed and aligned
- The use of unit economics to minimize the complexity of cloud costs
- Metric-driven cost optimization and gamification to keep teams aligned and focused on making every dollar count.

06

How IBM can help

No matter which FinOps phase you're focused on or how mature your practice is, IBM offers solutions that can grow with you. By offering advanced insights, planning and automation, IBM can help you maximize the value of your tech spend in essentially any cloud environment.

The IBM® Cloudability® solution provides a granular view into cloud expenses, offering tools that help organizations track, analyze and optimize their cloud spend. Its robust financial management framework helps businesses track spend by consumer, measure unit economics and enforce internal policy with automation. IBM Turbonomic® solution extends these capabilities into the operational domain, using AI-driven analytics to make real-time adjustments that balance performance needs with cost-efficiency.

IBM FinOps solutions can also help you overcome common FinOps blockers as follows:

FinOps blockers	IBM FinOps capabilities
Inconsistent tagging and costing across various cloud providers	View and allocate costs across a multicloud environment.
Inability to easily detect cloud waste, anomalies, or orphaned or idle resources	Reduce waste caused by overprovisioned and idle cloud resources.
Inability to easily compare multicloud discount programs	Maximize and automate coverage and utilization of commitments.
Inability to assign or reduce container costs in the cloud	Manage and optimize the cost of containerized infrastructure.
Inability to determine complete set of cost drivers or allocate shared costs accurately	Calculate the total cost of running products in the cloud and allocate those costs to the teams consuming them.
Inability to connect cost and revenue data holistically	Measure unit cost and mature to unit economics.
Reliance on overprovisioning to assure application performance	Unlock elasticity, scaling up or down based on needs while ensuring performance remains stable.
Inability to align performance and cost to business requirements	Automate optimization based on business outcomes.

IBM in action

How a multinational cosmetics company reduced its public cloud spend without compromising the end-user experience

Natura, the sixth largest direct sales company in the world, is known for offering beauty and personal care solutions in a conscious and responsible manner. Even though the company's existing system was fully functional, it needed a more proactive approach to resource allocation and cost management. With IBM Turbonomic software, the company improved efficiency and performance with AI-powered automation.

USD 260,000
Saved over USD 260,000 on public cloud spend over a 12-month period

>5,800
Executed over 5,800 automated resourcing actions over a 90-day period

IBM in action

When wise money management is a nonprofit's top priority

Cost avoidance plus the need to modernize its applications led the National Rural Electric Cooperative Association (NREC) to explore migrating to the cloud. When it realized it needed more visibility into cloud expenses, NREC turned to IBM for help. With the Cloudability tool, the organization fully allocated cloud spend and correlated it with business value while optimizing costs.

30%
reduction in infrastructure costs

70%–90%
reduction in cloud expenses

Conclusion

The discipline of FinOps is maturing to the point where each dollar of IT spend may be directly tied to some measurement of business value. This capability can be a huge competitive advantage, indicating when investments are paying off and when it’s time to change course.

That’s why it’s crucial to master the basics of visibility, along with a continuous focus on optimization, augmented by trusted, automated action. You must also consider the long-term role of FinOps within your organization and focus on how to build a culture that values the input of everyone involved in the process and delivers the capabilities they need.

And with the possibilities gen AI delivers—such as using NLP for pattern detection and robust reporting from disparate data—you’ll have access to insights and recommendations that wouldn’t have been possible before. Given the amount of information gen AI can provide, how you orchestrate your processes may be one of most critical factors to your overall success.

1. Tech spend: How will you pay for it? IBM Institute for Business Value, 26 September 2023.

2. 2024 State of the Cloud Report, Flexera, 2024.

3. State of FinOps 2024 Report, FinOps Foundation, 22 February 2024.

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
June 2024