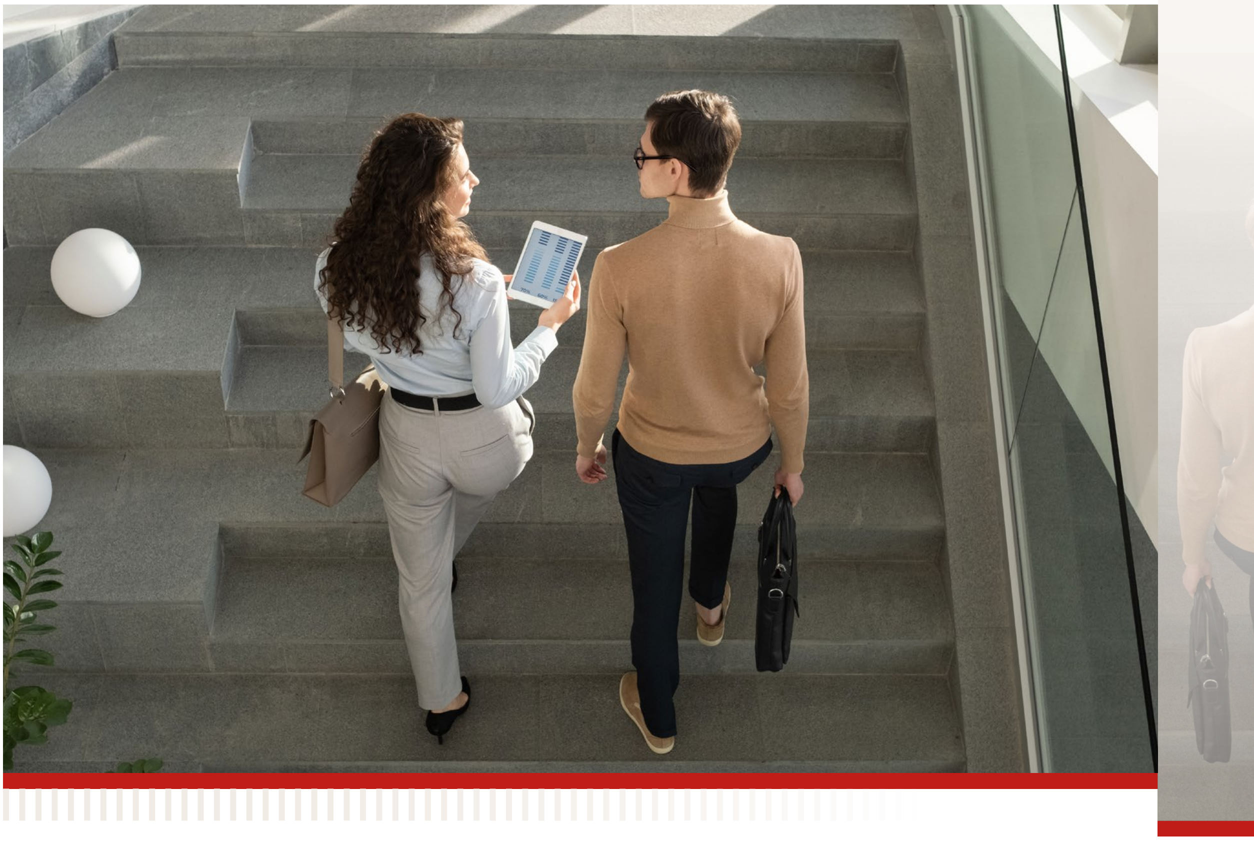
SPONSORED CONTENT | WHITE PAPER

Proactive, continuous, automated: why cloud cost automation isn’t a one-time fix

Managing cloud costs requires leaders to recognize the uniqueness of this challenge and design a robust plan that brings together key players and objectives.



CIO

SPONSORED BY <epam>

Analysts predict that cloud computing will unlock over $1 trillion of new value for companies by the end of the decade. These sources of value are familiar: cost reductions, instant scalability, improved resilience, digital transformation, more innovation, and accelerated product development.

But for many organizations, the cloud dream isn't delivering. Cloud cost optimization (CCO), for example, is a widespread challenge. In many cases, simply lifting and shifting infrastructure from data centers to the cloud leads to increased costs. Why? One reason: The very simplicity of cloud can easily lead to uncontrolled sprawl of resources. The impacts can include high costs and management complexity as well as security and compliance vulnerabilities. According to one study, enterprises typically estimate that 30% of their cloud spend is wasted.

Managing costs effectively is an essential part of making the transition to cloud computing. However, it is not unknown for enterprises to halt, or even reverse, their cloud migration efforts after encountering seemingly intractable cost management challenges.

In particular, senior executives who consider cost management one of their core competencies need to recognize that optimizing the cost of cloud computing is unlike most things they have done before. Success requires an understanding of rival architectures and platforms, a willingness to embrace systems design, and the ability to drive significant organizational change. When enterprises migrate to the cloud, management needs to shift accountability for costs decisively, from the central IT organization to those consuming resources, namely the product development teams developing new applications and services. On both sides of this transition, a significant cultural change is required.

The need for cost-aware architecture

Optimizing the cost of cloud consumption should be a continuous process, not a reactive one-time process that begins after an application has been deployed.

Cost optimization begins with cost visibility and accountability. Consider designing accounts and subscriptions to align with organizational units, and use resource tagging – attaching identifiers to cloud resources – to ensure that the right teams are billed for the cloud services they use.

It makes sense to build cost optimization into initial exercises in systems design. Using cloud service providers' guidance on the costs of specific configurations (such as cost estimators), architects can develop cost projections for senior IT and business managers at the very earliest stages of migration planning.

The architectural choices developers make can have a substantial effect on costs. One typical example involves deciding how to run a high-load system in the public cloud. One possibility is to use virtual machines sitting behind a load balancer. Another is to use lambdas (or functions) within a serverless environment. As Denis Kharlamov, cloud delivery director at EPAM Systems, argues, "Using lambdas seems to be the more cloud-native solution. But in a high-load system, this solution will cost you hundreds of times more than using simple virtual machines."

Developers need to consider costs as well as benefits. In the right context, the developer's natural preference for exploring novel solutions can produce remarkable feats of innovation. However, a preference for novelty often results in a failure to pursue solutions that can produce the desired results at a much lower cost.

As Krishna Challa, director of technology solutions at EPAM Systems, sees it, cost optimization is about asking questions: "Cost optimization involves engineering practices. For example, there are tools available, many of them open source, which help developers to estimate the cost of provisioning and scaling capacity in the cloud. But these tools — and the way in which they are used — need to be aligned with requirements such as load, performance, and resilience. In addition, engineers need to ask themselves: How well is the underlying code structured? Does the architecture take into account additional use case scenarios we may wish to develop in the future? For the performance we need, are we using the right service or are better options available?"

Proactive or reactive?

Cloud costs can be incurred unexpectedly for a variety of reasons. For example, it's quite easy to add more resources when the load increases — and then forget to release them as the load declines. New costs can also arise when applications are refactored or when net-new cloud services — such as machine learning — are deployed to generate additional business value.

Thus, it's important to take advantage of the cost management tools offered by major cloud vendors and by software vendors — but recognize their limits. As Challa puts it, "Tooling of this kind gives us visibility into where the spend is going and a granular breakdown. It gives us forecasts, and it identifies cost anomalies. But by itself, this won't solve a cost optimization problem."

These tools can facilitate cost analysis, forecasting, and budgeting. After all, cloud service providers have a vested interest in avoiding runaway cost crises. The tools they offer are worth understanding and exploiting.

Continual monitoring: building a process

Monitoring and alerts are sources of insight that enable ongoing discussions about cost strategy. IT organizations that proactively manage cloud costs typically set up a series of gated decision points during each application's life cycle. The first of these — a monthly or annual run rate projection for the service in question — should be included in the product development team's initial architecture proposals. Next, this projection becomes the basis for a budget around which alert triggers are defined (e.g., exceeding 75% of budget might trigger one type of alert, whereas exceeding 100% will trigger another alert). Developers, business managers, and those responsible for budgets need to review and agree on these thresholds.

Control mechanisms that define the consequences of cost overruns are needed, too. So, for example, it will make sense for most organizations to designate a series of core systems that continue operating 24/7 even if overall cloud costs rise above predefined limits. And when budget overruns look like a possibility, who should authorize them? As Challa says, these controls can either allow "multiple individuals to sign off on overruns of a specific size" or they can be implemented "more dictatorially" by a more limited number of executives.

The potential of automation and AI

Enterprises should automate as much of this process as possible. For Kharlamov, repeatable parameterized scripts are a "necessary foundation" for infrastructure provisioning and updates as well as ongoing cost optimization. At the simplest level, scheduling and switching off unused resources can be automated with scripts. The same approach can be extended to the entire cloud estate.

Once automated processes are harvesting the right data sets, understanding current consumption becomes possible on dashboards. The next step is to forecast. As Challa puts it, "Once a basic level of hygiene is in place around automation, then we can talk about using AI models to forecast and then optimize the future."

The rise of FinOps

It's not unusual for organizations to struggle with the costs of public cloud. However, moving beyond firefighting typically requires organizational transformation. The skills, culture, and organizational structures required to successfully manage the cloud differ radically from on-premises norms.

IT organizations accustomed to managing large capital expenditure (CapEx) budgets often find the shift to operational expenditure (OpEx) challenging. Likewise, development teams have traditionally managed their own people costs but can be less accustomed to managing the machine-based costs incurred by their code. The wider question is how these and other siloed corporate functions can work together.

This is the territory of FinOps, a discipline that enterprises increasingly use to put in place the organizational structures that encourage collaboration between IT operations, developers, finance, the compliance department, and representatives from the business.

Breaking down silos can be tough work. When it comes to culture and organization, leadership is crucial.

As Challa puts it, "There is occasionally friction between these siloed teams around protecting their own territory. They want to know who has control of what. People know that change is coming. But introducing that change becomes a lot easier when you have the right people involved in the right roles."

Rightsizing consumption during a crisis

Much of this paper has been devoted to the need to adopt a strategic approach to cost optimization. But if your organization finds itself in the middle of a cloud cost crisis, there are tried and tested ways to reduce consumption in the short term:

- **Shut down unused resources at night or on weekends, for example.**

- **Resize underutilized resources and configure autoscaling correctly.**

- **In multicloud environments especially, reduce the number and cost of outbound data transfers.**

- **Ensure that your organization's data is stored at the right tier, with the right capacity and appropriate cost.**

- **Investigate the potential of reserved compute capacity (up to 70% cheaper than pay-as-you-go), savings plans (discounts for longer-term spending commitments), and spot instances (inexpensive time-limited capacity).**

Finding a route out of cloud cost crises

Cloud technologies do not exist on a kind of island in the middle of the enterprise, isolated from surrounding organizational structures and cultures. Far from it. Just as the cloud is remaking infrastructure and applications, it is remaking organizations, too.

Cloud cost optimization is just one facet of an expanding requirement for operating models that also covers data privacy, industry standards, regulatory compliance, and security threats.

Cost optimization needs to become a continuous rather than an annual or monthly process, embedded in the entire life cycle of cloud development and migration, from strategy to execution. The requirements include standardized processes, pervasive automation, and increasing deployment of artificial intelligence and machine learning.

Most of all, however, cost optimization requires organizational and cultural change. Enterprises need to encourage their developers and wider product development teams to take responsibility for the cost of the cloud. In most enterprises, this change sets the scene for successful cloud cost optimization, which, in turn, opens the way to exploiting the full potential of public cloud.

About EPAM Systems

EPAM Systems Inc. (EPAM), a leading digital transformation services and product engineering company. Since 1993, the Company has leveraged its advanced software engineering heritage to become the foremost global digital transformation services provider – leading the industry in digital and physical product development and digital platform engineering services. Through its innovative strategy; integrated advisory, consulting, and design capabilities; and unique 'Engineering DNA,' EPAM's globally deployed hybrid teams help make the future real for clients and communities around the world by powering better enterprise, education and health platforms that connect people, optimize experiences and improve people's lives.

To learn more about our approach to using the cloud the native way, as it was intended, visit our website.◆