APPTIO
an IBM Company

eBook

FinOps: A New Approach to Cloud Financial Management

Table of Contents

Adopting and scaling FinOps

Executive summary

DevOps in public cloud has broken traditional infrastructure procurement — which is capital-heavy and slow-moving. With public cloud, procurement responsibility has been effectively outsourced to engineering. Engineers now spend company money at will and make financial decisions on cloud providers like AWS, GCP, and Azure at rapid speed. The result is engineers making daily decisions that affect the bottom line of their companies while finance teams struggle to keep up with the pace and granularity of spending. This change has necessitated the discipline of FinOps.

FinOps is the operating model for cloud financial management. FinOps enables a shift — a combination of systems, best practices, and culture — to increase an organization's ability to understand cloud costs and make trade-offs. In the same way that DevOps revolutionized software development by breaking down silos and increasing agility, FinOps increases the business value of cloud by bringing together technology, business, and finance professionals with a new set of processes — enabling cloud leaders to master the unit economics of cloud and drive competitive advantage.

Introduction

Public cloud is now widely adopted across businesses large and small, from traditional enterprises to the most innovative startups, with traction across all key verticals. The COVID-19 pandemic served to accelerate this trend as organizations seek to increase their operational agility, prioritize innovation, and look for potential cost savings. In a recent study, Gartner estimated that worldwide, public cloud spend is growing at 19.6% (CAGR), with annual spend expected to reach $661 billion by 2025 (vs. $340 billion in 2021). Additionally, Gartner forecasts that by 2025, "almost two-thirds (65.9%) of spending on application software will be directed toward cloud technologies".

Simply put, cloud is the new normal.

To fully get the benefits of this shift, organizations must adapt to this completely new way of procuring IT infrastructure — going from predictable upfront capital expenditure to variable consumption-based monthly bills and from procurement with strong internal gating to a world where developers (and robots!) provision resources at will. This has spurred the creation of a thriving new community and practice called FinOps, otherwise known as Cloud Financial Management.

FinOps is a combination of systems, best practices, and culture that increases an organization's ability to understand cloud costs and make trade-offs between speed, quality, and cost. In the same way that DevOps revolutionized software development by breaking down silos and increasing agility, FinOps maximizes the business value of cloud by bringing together technology, business, and finance professionals with a shared set of processes.

… almost two-thirds (65.9%) of spending on application software will be directed toward cloud technologies

Source: Gartner Press Release - Feb 22

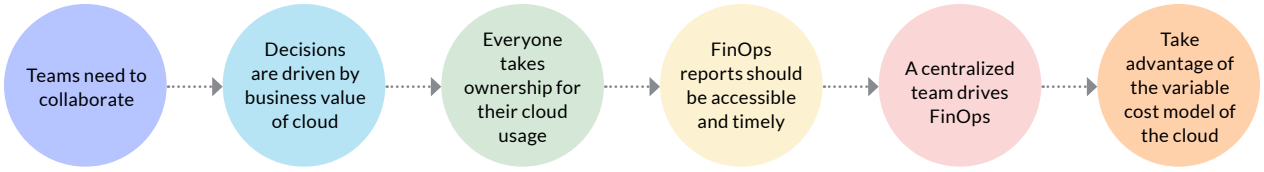
What is FinOps?

FinOps is a continuous, iterative journey that traditional enterprises and cloud-native organizations embark on as they continue to adopt cloud. The goal of FinOps is to balance cost, speed, and quality in order to gain cloud efficiencies and keep reinvesting in innovation.

The six principles of FinOps

FinOps is built around six core principles. FinOps practices that embrace these principles will be able to establish a self-governing, cost-conscious culture that promotes both cost accountability and business agility to better manage and optimize costs while maintaining the velocity and innovation benefits of cloud.

These principles were developed by the FinOps Foundation and validated by AWS:



Making sure that all your processes, tooling, and people align with these principles will help ensure the success of your FinOps practice. Let's look at them a little closer.

1 **Teams need to collaborate**

FinOps is about cultural change and breaking down the silos between teams that historically haven't worked together. In order to be effective, your FinOps practice needs to promote collaborative, productive conversations and actions.

2 **Decisions are driven by the business value of cloud**

Building a FinOps practice is about more than just reducing costs — it's about maximizing the impact of cloud. Sometimes that means spending more to drive innovation. The goal is to make sure you deliberately make the decision to increase spend instead of allowing spend to creep up from waste without creating business value. FinOps should always consider the business value of cloud with the goal of making completely informed decisions.

3 **Everyone takes ownership for their cloud usage**

The core idea of cloud billing is straightforward — if you use more, then you pay more. As a correlation, that means if you're responsible for cloud usage, then you're responsible for cloud costs. To keep costs under control, financial accountability must be spread to the edges of your organization, all the way to individual engineers and their teams.

4 **FinOps reports should be accessible and timely**

In the world of per-second compute resource billing and automated deployments, monthly or quarterly cost reporting isn't good enough. With cloud vendors reporting cost data in near real-time, your teams need self-serve access to this information to quickly understand the impact of their infrastructure decisions.

5 **A centralized team drives FinOps**

If you want to change your culture, you need someone to drive it forward. A central FinOps team drives best practices through standardization, education, and cheerleading. That same team can centralize rate optimizations — through commitment-based discounts and negotiating enterprise discounts — to take full advantage of them while empowering the rest of the organization to action usage optimizations.

6 **Take advantage of the variable cost model of cloud**

While adapting to the variable cost model of cloud comes with its challenges, it also has its advantages. Instead of heavy capacity planning cycles where teams are constantly asking "What will we need to cover future demand?" you can adjust your provisioned infrastructure dynamically to meet fluctuating demand. Not only is this more efficient as there is no longer a need to maintain capacity for peak load periods — with idle capacity between peaks — you can confidently adapt to changing conditions as needed.

Cross-company collaboration is the key

The goal of FinOps isn't necessarily to spend less — rather, to make sure a company's cloud spend is optimized and that the company is getting an adequate return on the investment made.

Achieving those goals requires fluid, cross-company communication across multiple teams. Adopting cloud can drive increased agility, innovation, decentralized decision-making, and fast adaptation to change. These qualities must be mirrored in how a company makes financial decisions about its cloud infrastructure.

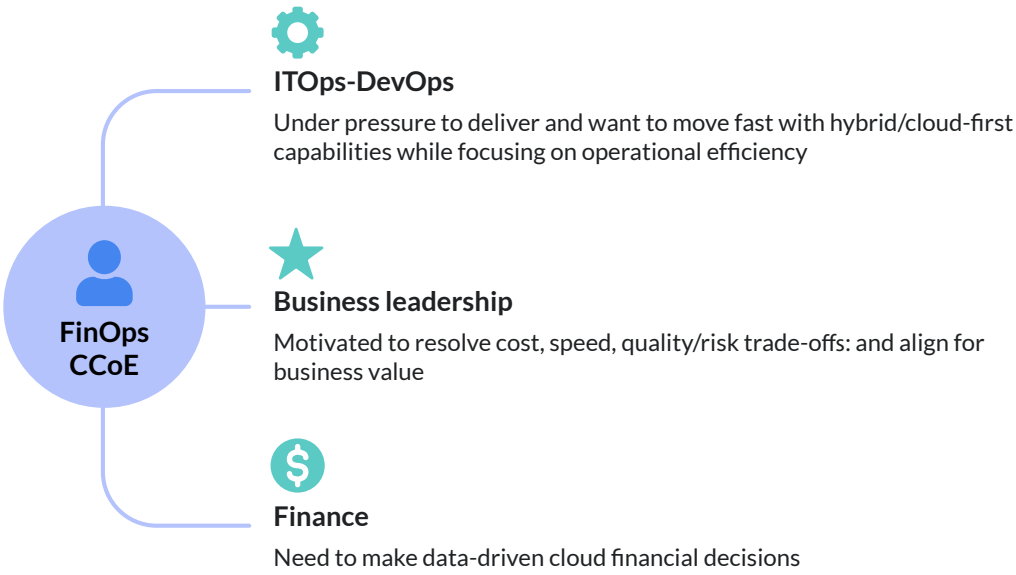
Driving this need is the shift from capital-heavy, fixed spending for infrastructure in the data center to the consumption-based, variable spending model of public cloud. Variable spending gives individual teams the ability to make daily decisions about whether provisioned resources are producing sufficient business value and adjust those resources for an immediate financial impact. Teams can measure the impact of that spending and iterate quickly in one of three directions: maximize the quality of their offering, improve speed of delivery, or drive down costs — also known as good, fast, or cheap.

In the same way that DevOps revolutionized software development by breaking down silos and increasing agility, FinOps increases the business value of cloud by breaking down additional silos — particularly into the finance team — and providing a shared set of cross-function processes.

Organizations either establish a dedicated FinOps team or create a focus group within the Cloud Center of Excellence (CCoE) that meets regularly to discuss choices about cloud infrastructure. A key goal for this group is to make sure everyone understands the interplay between the actual infrastructure, infrastructure costs, and business goals. Staff from finance can take on a financial planning and advisory role. Management can give its perspective on what exactly should be optimized in terms of cost, speed, and quality. Engineering can contribute by explaining what cloud resources they need to build the applications and features management has identified as adding value to the business.

Cross-functional FinOps team drives best practices

FinOps drives best practices and helps people make sensible decisions about real-world trade-offs. Finance teams are still focused on costs and allocation, but now they're partners with the technology and business teams. They can shift from CapEx reporting to OpEx forecasting and work with those who understand what the drivers of cloud spend are even when they are consuming thousands of SKUs. For example, the VP of finance can sit down with a DevOps team and have actual conversations about the trade-offs of going over budget compared to the value that increased spending brings by enabling the company to ship key features faster, which can ultimately mean more revenue. Together, these stakeholders can come to a decision about the best approach to take.



Involving the stakeholders

Because infrastructure procurement shifts from static, siloed processes to fluid, agile ones, members of a FinOps team must move outside their comfort zone and learn something about each other's disciplines. Remember, it's all about balancing cost, time, and quality — and everyone has a say in how to do that.

Finance learns about cloud

Before FinOps, a finance team might have looked at spending monthly or even quarterly to see, for example, the total cost for compute resources. When dealing with public cloud, a finance team working in isolation can quickly become overwhelmed with the explosion of SKUs and services that are characteristic of cloud vendors.

With FinOps, finance meets frequently with the technology team to understand how those cloud resources are used, what projects might be coming up, and what resources they expect to need. Instead of examining costs

in terms of months or years, the finance members shift the time frame to resolutions of days and hours, which reflects how cloud vendors calculate charges.

With the knowledge they get from the business and engineering members, stakeholders from finance can take specific actions, such as buying savings plans or renegotiating terms with a cloud provider.

Engineering learns to consider costs

A crucial point for engineering is to understand the impact of the shift from CapEx to OpEx and the consumption-based spend model of cloud. The cloud has given engineers, who can spin up their own compute resources in a moment, the power to affect OpEx spending with their daily decision-making, which in turn affects the company's bottom line. If OpEx spending is treated as a cost-of-goods expense (COGS), engineering actions to procure cloud resources hit the margins of a publicly listed company and can ultimately affect its stock price. It's critical to get a handle on how these costs are reported and allocated.

Just as finance teams have to broaden their scope to understand what makes up cloud infrastructure, engineers, whether in operations or development, also need to understand how their infrastructure choices affect the company's finances. Being a good engineer isn't just about understanding the tech anymore; engineers must understand how technological choices have a financial impact and consider cost as a new efficiency metric when they write code and deploy infrastructure.

Management sets the priorities

Management needs to set business priorities so that other FinOps team members can start to think about the trade-offs between cost, time, and quality. Management also needs to understand what choices are available to finance and engineering to make the most informed decisions that will best support the business. The Iron Triangle is management's guide. They take a broad view and make the decisions on whether to optimize for time, quality, or cost for each product.

The language of FinOps

Communication among FinOps team members is essential to effectively optimize for speed, cost, and quality. Because the team includes people from different disciplines, everyone needs to share a common vocabulary for describing cloud spend. This is the language of FinOps.

In terms of optimization, the two most important terms are usage optimization and financial optimization. Usage optimization refers to activities where you turn off a resource or reduce its size to a cheaper option to decrease cost.

Usage optimization is about using less. Financial optimization, on the other hand, is about reducing the rates you pay for what you are using. For example, you might optimize costs with savings plans (on Amazon or Azure) or committed use discounts (on GCP). Once every member of the FinOps team understands what usage optimization and financial optimization are, they can discuss the implications for their cloud bill.

At its simplest, a bill can be summarized as: Spend = Usage x Rate. This equation implies that usage optimization, which relates to how much you use, and financial optimization, which relates to hourly rates, are the two levers available to reduce the amount the company spends.

Measuring usage

When looking at usage optimization, it's important to understand how usage is measured. Usage isn't simply a count of the number of services a company uses. Each cloud service uses different metrics to measure usage. To understand the cloud bill, it's important to understand how the cloud provider charges for each of them.

For example, on AWS, for EC2 instances, you're charged per hour of usage. What matters is time or how long you run each instance. On GCP, the storage costs for BigQuery are charged by GB-month, so you're charged for a combination of the size and duration of storage. On Azure, any data coming out of a VM to a non-Azure location is called egress traffic and charged per GB. The first 100 GB of data are free each month. Above that limit, you're going to have to estimate your egress traffic. When you estimate costs, make sure you understand how each resource type is charged by the cloud provider to avoid surprises.

Usage optimization activities

Two of the most common usage optimization activities are rightsizing and automated scheduling. Rightsizing's goal is to provision resources that are well matched to the needs of the underlying workloads. The infrastructure is neither over- nor under-provisioned. The resources, such as VMs or managed databases, have enough capacity to get the job done without "clipping" but not so much capacity that resources get wasted. Likewise, for storage-based services, performance attributes such as throughput and IOPS capacity should be matched to the requirements of the workload while avoiding provisioning them in excess. Remember that you are paying for what you provision, not what you actually use.

Automated scheduling takes advantage of cloud's elasticity by programmatically handling repetitive or hygiene tasks, such as shutting down resources that aren't actively used. For example, a company might write a script that turns off VM instances that aren't used over the weekend and then starts them again for Monday morning. They might also regularly run a script that snapshots and then deletes orphaned block storage volumes. Since these volumes aren't attached to a VM, there's a good chance they're racking up hourly charges while doing nothing.

Making usage-optimization decisions

Making usage-optimization decisions often involves the whole FinOps team. Engineering needs to be involved because they'll make the actual changes to the infrastructure and understand the implications of those changes in terms of performance and potential impact on customers. Management is a stakeholder in making sure that business goals are met. Finance is there to track, forecast, and monitor how the decisions that engineering and management make impact costs.

Measure everything

The FinOps team should use metrics to make sure every usage-optimization and financial-optimization activity is paying off. Are the savings plans you have in place being used? Are they meeting your savings goals? Are the VM instances you're using for a particular project the right size? Are they delivering enough performance to ensure customer satisfaction? Is the CPU usage well below what the instance can provide?

Financial optimization activities

Financial optimization decisions can be made by the finance members of the FinOps team and are often informed by usage-optimization decisions. For example, Finance can purchase commitment-based discounts like savings plans on Amazon or Azure and committed use discounts (CUDs) on GCP. If your organization has a large annual spend, you can also choose to directly negotiate enterprise discounts with each vendor. The finance and procurement teams can use their centralized visibility to drive higher coverage of these commitments across their entire spend and to negotiate better terms.

All of these metrics need a target. A measurement without a target is just data, meaning you won't know if you're on track or if you need to change something. For example, a target on Amazon might be to make sure that all savings plans have 90% utilization. To make sure that target is being met, you can track the actual utilization for every subscription you have on Amazon. Measure the number of hours that each subscription is getting applied versus hours when it is not. If any one of those subscriptions falls below 90% utilization, you can consider certain options to increase its usage and approach further commitments with caution.

Metrics with targets make it possible to have objective conversations with teams that aren't meeting those targets to understand why. Members of the FinOps team can work with them to decide on the best actions for evaluating those targets and, if they're realistic, reaching them.

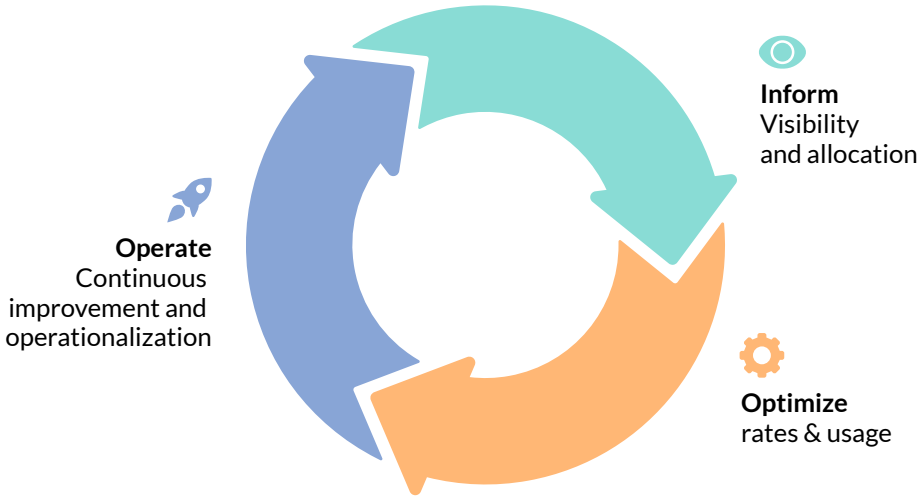
The FinOps lifecycle

The implementation of FinOps takes place across three phases, with each phase guided by the six core principles. Those phases are Inform, Optimize, and Operate. When FinOps is used correctly, organizations cycle through these three phases constantly, achieving higher and higher levels of success.

The Inform phase gives stakeholders visibility into the cloud bill and allocates costs back to the business groups responsible. This encourages financial accountability by showing teams what they're spending and why. This phase enables individuals to see the impact of their actions on the bill.

The Optimize phase is focused on making changes to cloud usage and how you pay for cloud to improve overall cost efficiency. This includes activities like rightsizing, tuning storage access frequency, and improving commitment coverage. Goals are set upon the optimizations identified, which align with each team's area of focus.

The Operate phase defines processes that enable the goals of technology, finance, and business to be achieved. Automation can be deployed to enable these processes to be performed in a reliable and repeatable manner.



FinOps Phases

The lifecycle is inherently a loop. The most successful companies take a crawl, walk, run approach and get a little better each time they go through it. Let's dig a little further into the activities that go along with each phase.

Mapping activities to the FinOps phases

Inform			Optimize		Operate
Showback or chargeback costs	**Visibility and awareness**	**Accountability**	**Optimize usage**	**Optimize rate**	**Continuous improvement**
Amortize and allocate commitments	Daily KPI updates	Trend and variance analysis	Idle resource detection	Commitment-based discounts	Evaluate metrics and processes
Apply custom pricing and FX	Dashboards and analytics	Budget and forecasts	Rightsizing	Enterprise discounts	Automate FinOps
Map spend back to the business	Cost event detection	Benchmarking	Automated scheduling	Spot markets	Unit economics

Inform

Showback or chargeback costs

Whether you are showing back or charging back cloud costs — the only material difference between these two processes is that charging back affects the profit & loss statement — the core requirements are the same: the need to establish the true cost of running cloud at a line-item level and then allocating these line items back to the business groups responsible. To establish the true cost of running cloud, it's necessary to allocate the commitments — which often have an upfront component — to the resources and teams that consume them.

Many organizations will agree to enterprise discounts with their vendor — typically a flat discount which may or may not appear in the detailed billing. An important decision for IT finance is whether to pass these discounts on or hold them back for other purposes.

Finally, another complication in establishing the true cost of running cloud can be handling foreign exchange rates, either because operations are global or bills are delivered in a foreign currency. Having established the true cost of running cloud, the FinOps team is well-placed to roll out an allocation strategy, mapping the costs back to the business based on billing attributes such as resource tags and accounts.

Visibility and awareness

With costs accurately and reliably allocated, the FinOps team is well positioned to make stakeholders across the business aware of their, and where applicable, the organizations spend on a frequent, if not daily, basis. One of the most effective ways to achieve this is to push daily summaries out to a broad audience via email or collaboration tool and include crucial information, such as the accrued spend for the current month and how it compares to the previous month.

The next layer of detail for stakeholders needing to answer questions about this spend is typically self-serve dashboards and analytics, with users able to craft reports and visualizations filtered to relevant data and highlight cost factors important to them. As the FinOps practice matures, proactive alerting can play a bigger role, with AI tooling able to detect unusual spending patterns and other cost events so that stakeholders can be notified in near real-time.

Accountability

The ultimate goal of the Inform phase is to hold individuals and teams accountable for their contribution to the organization's cloud bill. While this relies on successful allocation and shared visibility, there are often extra steps needed to drive accountability and change behavior. Nothing achieves this like having official budgets that are tracked month-to-month. Another motivating option employed by mature FinOps teams is to roll out benchmarking, allowing performance comparisons — such as waste percentages and commitment coverage — between internal teams and against industry cohorts.

Optimize

Usage

Strong implementation of financial accountability during the Inform phase positions organizations well to succeed at optimizing usage. Since the task of optimizing usage falls on the shoulders of engineers within delivery teams, the key focus for the FinOps team is getting actionable insights into the hands of these engineers. The highest priority should be identifying resources that are simply going unused but for which you are paying by the hour. Next in line are resources that are oversized or in the wrong shape and therefore leading to excessively high hourly rates. In this case, getting recommendations for alternative resource types or configurations backed by supporting utilization and savings information is crucial.

Rates

A primary job for FinOps practitioners and the IT Finance team in the context of managing cloud spend is using their expertise to lower the hourly rates paid for the deployed resources. While it's common for organizations to prioritize optimizing their usage before focusing on rates, it is worth noting that these are separate concerns with levers that can be pulled with some independence. Both commitment-based discounts and enterprise discounts require detailed analysis of historical usage patterns so that there can be a level of confidence when forecasting future consumption. Therefore, it's important to get timely and granular usage data that is categorized in business-meaningful ways to those responsible for negotiating enterprise discounts with the cloud vendors and for attaining commitment coverage levels.

Operate

Continuous improvement

The Operate phase is an opportunity to operationalize FinOps processes, automating manual steps where possible and evaluating whether cloud initiatives are delivering the expected business value. Across the FinOps lifecycle, there are many activities that can benefit from automation. These include tasks such as automating the chargeback of costs, onboarding users to FinOps tooling, and pushing out recommendations to the relevant stakeholders. This may be achieved by using the public APIs of your FinOps tooling or inbuilt workflow automation.

It's important to regularly evaluate your FinOps metrics, both to measure your progress against them and to consider whether they are the right metrics. Prime amongst these are unit cost metrics, bringing to light the cost it took to deliver each dollar of sales, each customer transaction, or 1000 website visits.

Implementing key FinOps practices

Allocating costs back to the business

As discussed previously, cost allocation lies at the heart of the Inform phase of the FinOps lifecycle. Companies who can successfully allocate costs to the business groups responsible set themselves up for FinOps success, while those who don't will invariably face an uphill battle. While every organization has different needs and will structure their cloud resources accordingly, it's worth spending the time to devise and institutionalize a deliberate cost-allocation strategy. In the upcoming sections, we will cover the key elements of such a strategy: the foundational hosting constructs, resource tags, and the process of applying business rules.

Hosting constructs - accounts, subscriptions, and projects

The first logical way to organize your cloud spend is through the use of accounts (that is, if you're using Amazon Web Services; If you're using Microsoft Azure, the equivalent is called subscriptions; if you're using the Google Cloud Platform, the equivalent is projects). Cloud providers allow you to create as many of these as you need. Whenever anyone provisions a cloud resource, such as a VM or volume, they must decide which account is most suitable for hosting it. This enables you to define an overarching structure for your cloud environment centered around logical conventions, naming and grouping accounts, subscriptions, or projects for different areas of your business (for example, by business unit).

In AWS, you can build this structure around one or a handful of payer accounts (now known as "management accounts"), with each backed by numerous member accounts (where the usage happens). For example, you could create a management account for each of your main business units. Then, you could create individual member accounts that are dedicated to product areas within the business unit. Enterprises are increasingly shifting to a single management account model to maximize reserved instance and savings plan benefits. Both Microsoft Azure and Google Cloud Platform provide similar grouping capabilities through subscriptions and projects, respectively.

It is worth noting that every resource, by necessity, belongs to an account, and therefore it's the most foolproof way of assigning top-level ownership. But if you need to capture more detailed information about what is driving your cloud costs, then you'll need to also consider leveraging resource tags.

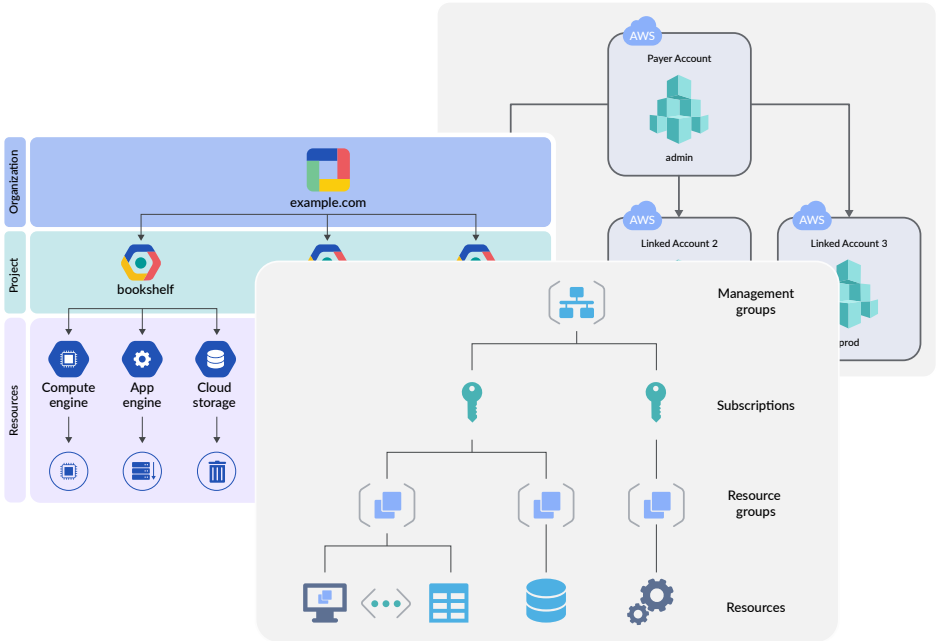


Figure: *Hosting constructs of GCP, Azure, and AWS*

Resource tags

Resource tags are like labels you put on physical assets. They consist of a customer-defined key-value pair. Tags can provide crucial information about any cloud resource, such as its name, environment, service, application, department, division, and much more. You can consider tags an important type of metadata that allows you to apply detailed information to the resources (i.e., compute, storage, and network services) your organization is consuming and, in turn, drive cost management.

Using tags, you can be as granular as you want in describing your resources. For instance, AWS allows 50 user-created tags to be attached to each resource, as does Microsoft Azure. GCP calls these labels instead of tags and allows up to 64 per resource.

But be aware. Tagging is case-sensitive, and spelling matters. And because you have the ability to attach so many attributes to each resource, it's easy to get carried away with tagging if you're not careful. In addition, not every resource type supports tags, so be sure to check your provider's limitations.

Your best approach is to start small. Use a tagging strategy that is simple and covers the most important reporting items, and then expand it over time if necessary.

Once your tagging and account strategy is in place, you'll be able to leverage attributes within detailed billing to allocate costs, then use showback or chargeback processes to hold business groups accountable.

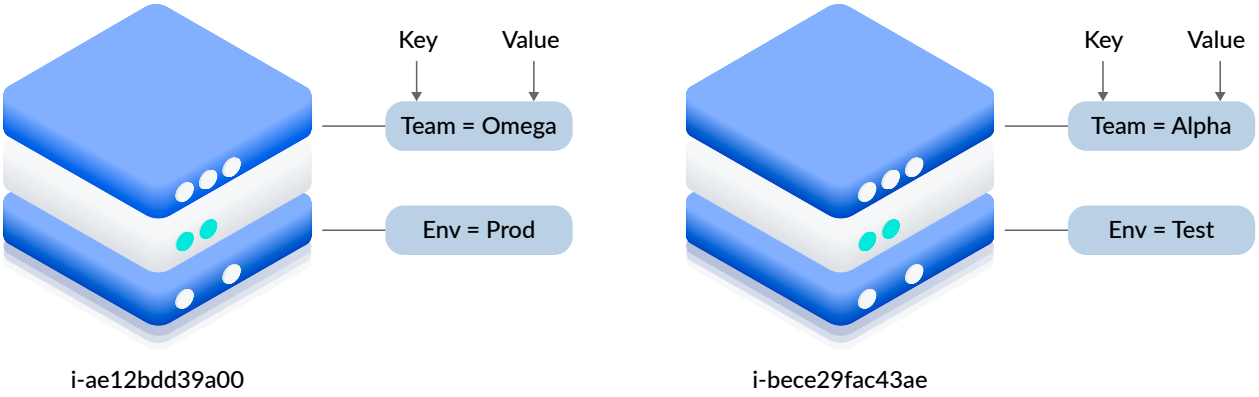


Figure: *How tags are applied to individual resources*

Applying business rules

Once you have invested in your foundations, and preferably institutionalized the agreed-upon patterns through build pipelines, you will be well-placed to implement processes to achieve a full chargeback or showback. This usually involves applying business rules to these foundations such that you establish official reporting definitions that are less aligned to your underlying infrastructure but more aligned to your business concepts. Broadly speaking, there are two reasons you need to build this additional layer.

The first scenario is due to requiring multiple billing attributes to assign ownership. Here are two common examples:

- Mixed-ownership model: An organization can have some accounts that are shared across teams and some that are team-owned. In this case, you need to establish an allocation rule that combines both tags and accounts.
- Account fallback model: Resource tags are the primary allocation mechanism. However, it is extremely rare to see every item tagged correctly, so you place an account-ownership mapping as a fallback. Again, your allocation rule is based around tags and accounts.

The second scenario involves taking the opportunity to greatly enrich the foundational data through mapping completely new business layers on top. Even though your base data may not explicitly have information such as the project ownership of a particular service or what type of expense an item is, you can create mapping rules to surface this detail. The most powerful examples we've seen recently involve connecting cloud financial management processes with enterprise IT service management (ITSM) workflows. This normally manifests itself with every cloud resource being tagged with a unique identifier for the service it backs. From there, as shown in the diagram below, you can map out the entire organization as it appears at any point in time.

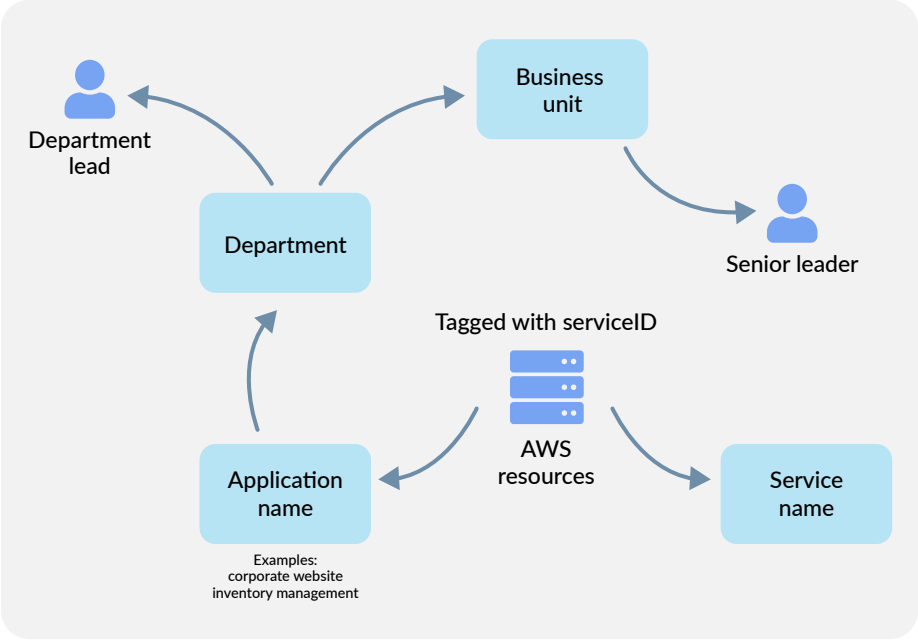


Figure: Illustration of mapping business dimensions from service ID

For more information on applying business rules to billing attributes to implement a chargeback of cloud costs, check out IBM Cloudability's business mapping.

Rightsizing cloud resources for maximum savings

For the usage component of the Optimize FinOps phase, rightsizing is the most prominent activity. Rightsizing refers to the processes and techniques for making intelligent trade-offs between performance and cost across cloud footprints. It requires an understanding of what your current cloud resources are costing you and to what degree each is being utilized (across metrics such as CPU and memory that must be gathered from monitoring solutions) and identifying alternative options (such as reduced size or different type) that will meet the needs of the underlying workload.

A simple rightsizing scenario might occur where an engineer provisions a large VM (hence high hourly rates) for a task that requires minimal CPU and memory. By rightsizing to a smaller VM, costs can be reduced with minimal performance risk. A more nuanced rightsizing scenario occurs where a provisioned resource has the wrong "shape". For example, a VM may be provisioned with a memory-optimized type even though the workload itself is compute-constrained. Moving to a well-fitting compute-optimized type would reduce the cost and potentially improve performance. There are many different rightsizing scenarios that can occur across many types of resources, including VMs, managed databases, and storage devices.

Due to the risks involved — the wrong change can negatively impact customers, after all — and challenges presented by the sheer scale of cloud — a typical scenario having tens of thousands of cloud resources running concurrently — consistently achieving rightsizing goals demands specialist tools and dedicated strategies. Here are some important considerations to help guide your rightsizing efforts:

- There is no such thing as the perfect VM or DB type; therefore, always consider multiple options.
- Evaluate historical resource consumption for all utilization metrics, especially memory.
- Consider peak utilization periods and avoid "clipping" where possible.
- Review possible VM or DB types across instance families to get the right shape.
- Review provisioned IOPS and throughput for storage devices, a common location of addressable waste.

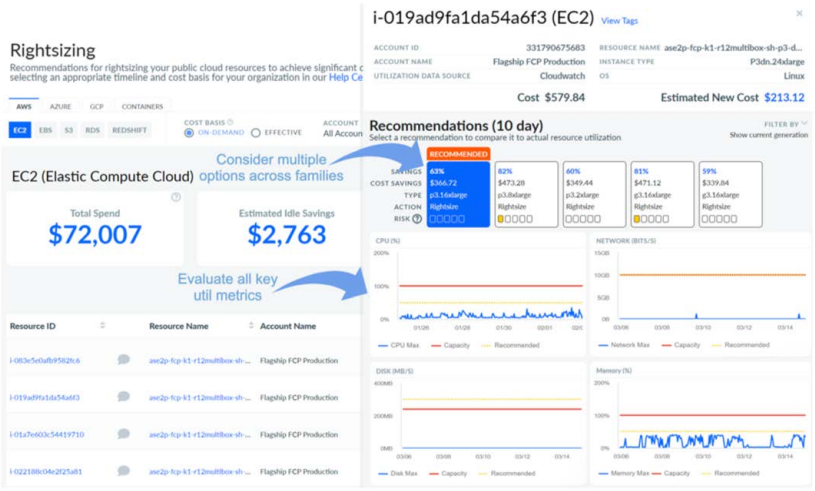


Figure: IBM Cloudability rightsizing example for AWS EC2 instance

For additional information on this topic, check out IBM Cloudability's rightsizing feature.

Reducing rates with commitment-based discounts

For the rates component of the Optimize FinOps phase, commitment-based discounts offer the largest opportunity to lower hourly rates, albeit requiring ongoing management to maximize benefits and minimize waste. Whether the instrument is a reserved instance (AWS, Azure), savings plan (AWS, Azure), or committed use discount (GCP), the overarching concept is the same — the vendors offer lower hourly rates for committing to specific types of usage for one- or three-year terms. You can think of these instruments as discount coupons; you purchase a set of coupons that get applied to your usage every hour. Every hour, the cloud vendor resolves which of your usage qualifies for the coupons and decides which specific resource usage will get the benefit (important: you can't purchase them for specific resources; instead, they "float around" based on the vendor's algorithms). Instead of being charged full on-demand fees — say, $1.00 for an hour — the usage will receive the discounted rate — say, $0.70.RIs are typically purchased based on three attributes: location, operating system, and VM type/size. For example, for AWS, you could choose to purchase 10 c7g.large EC2 RIs in US-East-1 for Linux. Savings plans are the latest commitment offering and superior to RIs in that individual commitments have a higher level of abstraction, covering a broader set of underlying usage. The most obvious difference is that savings plans are purchased in dollars ($) rather than instance count.

When purchasing a commitment, you choose a length and payment model (all upfront, partial upfront, no upfront), with the savings rate dependent on these choices. If any of these commitments go unused for a given hour, i.e., there wasn't adequate usage to consume them, then they become waste.

The importance of targeting an appropriate commitment level can be illustrated by the diagram below, which simplifies how commitments are consumed. In the histogram, each lettered block represents an individual VM that was running during a 10-hour period. In this example, VMs A and B ran for every hour, while all other instances ran for only part of this period.

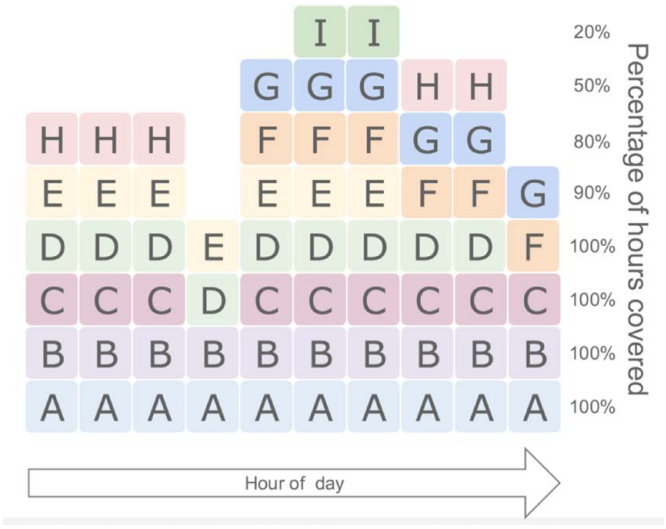


Figure: *Histogram of hypothetical VM usage*

Let's say that this 10-hour block is repeated for one year. We can see that if we purchased four applicable RIs, they would each have 100% utilization; a fifth RI would have 90% utilization; a sixth RI, 80%; and so on. While it could make sense to purchase the sixth RI — since its savings rate may be high enough to compensate for its unutilized hours — it's unlikely that you'd purchase a seventh RI, as it would only have 50% utilization.

Given the complexity around savings rates and the need to analyze hourly data across a wide combination of usage, we advise using a purpose-built recommendation engine that allows you to create a plan matched to your organization.

Extending FinOps to containerized workloads

As you begin to deploy more advanced technologies like containers, cloud cost allocation becomes more challenging.

Containerization makes it easier to package and run software applications in different computing environments. Running containers at scale is done by leveraging an orchestration tool, typically Kubernetes (K8s). With operational benefits around agility and resource efficiency, it's no surprise that K8s deployments are steadily growing as a percentage of public cloud spend.

From a cost allocation perspective, it's worth noting that each K8s cluster is usually a multi-tenant entity, with underlying virtual machines and volumes being shared across applications and teams. This is why some people refer to these deployments as a cloud-inside-a-cloud and describe a "black-box" problem in attempting to understand who is responsible for consuming the associated cloud resources.

The cloud vendors offer no native billing construct to surface the cost of each cluster or means to split their cost up so they can be allocated back to the business. It's highly recommended to take advantage of specialist solutions, such as IBM Cloudability, that can automatically map resources to their cluster, calculate total cluster cost, and then apply rules based on usage patterns to split these costs up and allocate them out via K8s constructs such as namespace and labels.

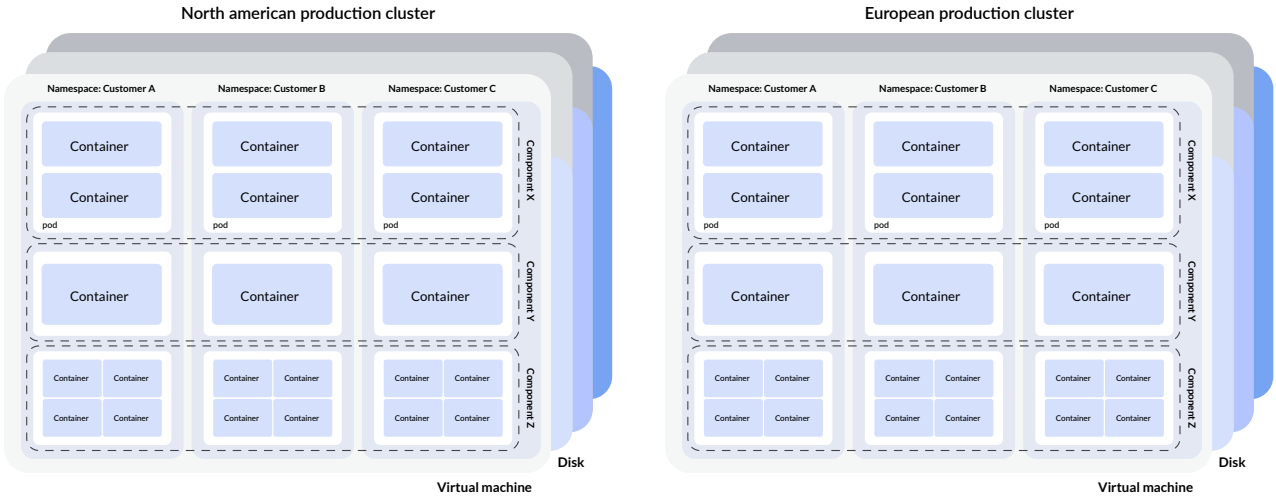


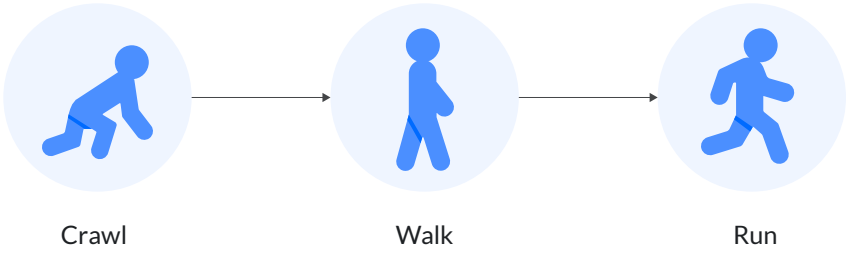
Figure: *Visual representation of Kubernetes clusters and their internal constructs*

Accelerate your FinOps journey today

As the adoption of public cloud continues to grow and revolutionize the procurement of IT infrastructure, we can expect FinOps to grow in prominence and become ever more critical to the success of an organization's cloud investment. Whether your role is technology, finance, or business-focused, and regardless of whether your company's cloud spend is large or small right now, it is likely that you can contribute to accelerating FinOps at your workplace. If you are struggling to know where to start, we recommend you consider the "Crawl, Walk, Run" approach as set out within the FinOps maturity model.

Another option is to pick one or more of the activities that we described in the FinOps lifecycle section of this book. For example, dedicating some time to identifying idle resources for remediation can lead to immediate wins and help build momentum.

And, of course, as leaders within the FinOps space, we at Apptio, an IBM company, are here to help you today with a rich set of relevant online resources, specialized tooling, and professional services.



FinOps maturity model

For more information

To learn more, contact your IBM Business Partner:

COMPANY NAME - VALUE
PHONE - VALUE | EMAIL - VALUE
WEBSITE - VALUE