Maximizing Your AWS Cost Savings

How to get the most out of your AWS spend while keeping costs under control

Published September 2024

Table of contents

Introduction

With 85% of companies planning to be “cloud-first” by 2025, as well as 75% of tech leaders saying that they’re building all new products and features in the cloud, cloud costs are almost certain to be the primary driver of IT spend over the course of the next decade.

Given the share of overall business spending that cloud services are due to take up, getting the most out of that investment is paramount. And yet, cloud inefficiencies are often sending those costs even higher, with up to one-third of cloud spend ultimately getting wasted.

Given that, it’s no wonder that cost optimization is at the forefront of every cloud practitioner’s mind.

Companies around the world are instituting FinOps practices to help manage and control this spend, with the goal of ensuring that their cloud budget is being put toward services that make the most sense for the long-term health of the business.

Yet adopting a wide-scale FinOps practice that involves key stakeholders across the organization is a long-term process that requires buy-in at all levels. The fastest means for building support and momentum is by producing quick wins with tangible results. This can build credibility within the wider organization, while creating a foundation that supports future steps.

One of the best ways to do this is by focusing on cost optimization.

This ebook will go over some of the areas within your AWS environment that may be the source of some cloud inefficiencies, and make optimization recommendations for how you can lower your monthly cloud bill without compromising any of your ongoing endeavors.

Best Practices for Ongoing Cost Optimization

The unfortunate truth is that your cost optimization work is never really finished. However, putting a process in place for conducting regular reviews or audits of your AWS environment should go a long way towards ensuring that your infrastructure is operating efficiently.

AWS Well-Architected Framework

The DoiT services team conducts cost optimization audits with customers as part of regular Well-Architected reviews. The AWS Well-Architected Framework provides best practices and guidelines for designing secure, high-performing, resilient, and efficient infrastructure for applications. It's comprised of six different pillars:

- Operational excellence
- Security
- Reliability
- Performance efficiency
- Cost optimization
- Sustainability

You should familiarize yourself with the entire framework to get the most out of your environment, and the cost optimization pillar can serve as a starting point for how you want to conduct your own audit.

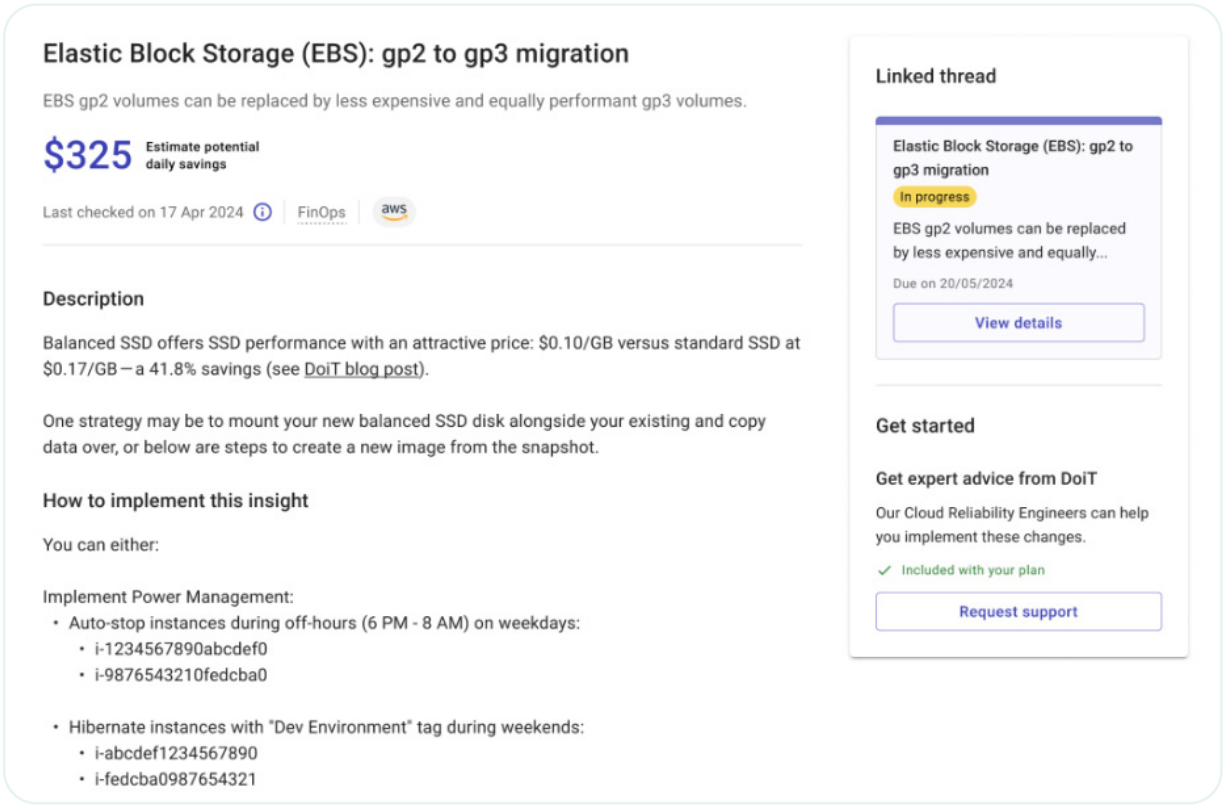
Assess skills and tools

Before conducting a cost optimization audit, you want to make sure that you have the proper tools at your disposal to do so. AWS provides native tooling that can get you at least part of the way there such as AWS Cost Explorer, AWS Budgets, and AWS Trusted Advisor, which can all be utilized for free at the lowest level, and but do have charges associated with some of their more advanced features. You should implement and utilize the free versions of these tools as a foundation, and then evaluate whether you want to pay for the advanced functionality or explore third-party SaaS tools to supplement your efforts.

It's also important to do an internal skill assessment to determine whether you have the necessary resources to conduct a cost optimization audit on your own, or whether you'd be better served by working with a partner who can not only bring new skills to the table that your team may not have, but also do a lot of the heavy lifting of the audit itself.

One thing to keep in mind is that the level of expertise needed depends on the complexity of your AWS infrastructure. For small to medium-sized businesses with straightforward setups, a basic understanding of AWS cost optimization principles and using AWS Cost Explorer may be sufficient. However, for large and complex environments, it's beneficial to have certified AWS professionals who can perform in-depth cost optimization analysis and implement complex cost-saving strategies using more advanced software solutions.

DoiT is able to provide this level of expertise as part of the DoiT Cloud Solve offering, which, among other things, includes in-depth cost optimization audits that can uncover many different savings opportunities that you may not be aware of. These in-depth audits can also be supplemented with DoiT Insights within DoiT Cloud Navigator, which continuously monitor for cost optimization opportunities and alert you whenever they arise. From there, you can take action on them by creating and tracking engineering tickets in Jira via DoiT Threads.



DoiT Insight with a linked Thread

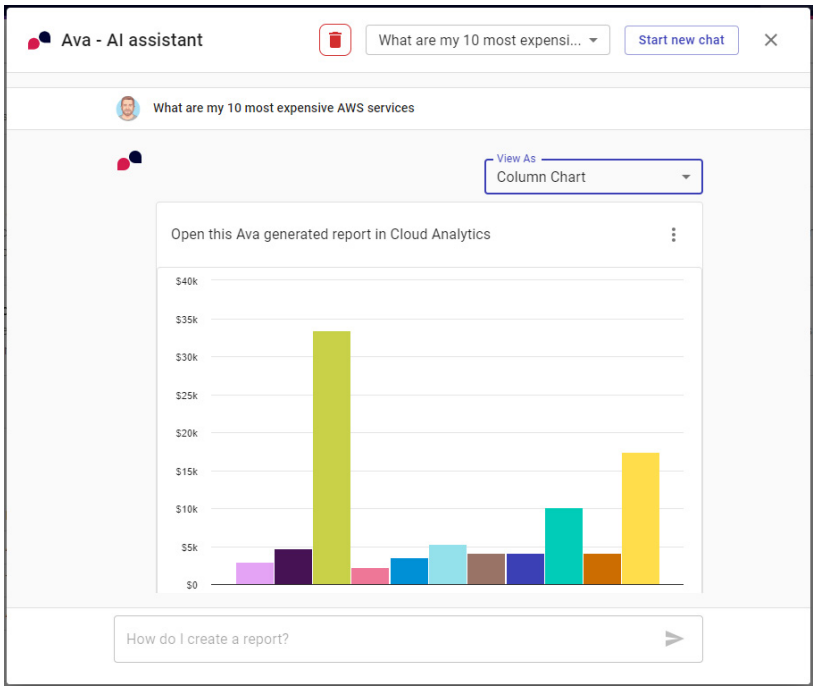
Frequency

You also must decide just how often you want to review your environment. Doing a full Well-Architected Review at least once a year or after any major infrastructure changes is strongly recommended, and if your environment is undergoing regular changes, then doing it every six months is advisable. You can also supplement this full-scale review with monthly audits that focus on your key cost drivers and any anomalous spend that has recently occurred. This will help ensure that your AWS infrastructure is consistently optimized, and hopefully reduce the amount of time-intensive work you have to do after your Well-Architected review.

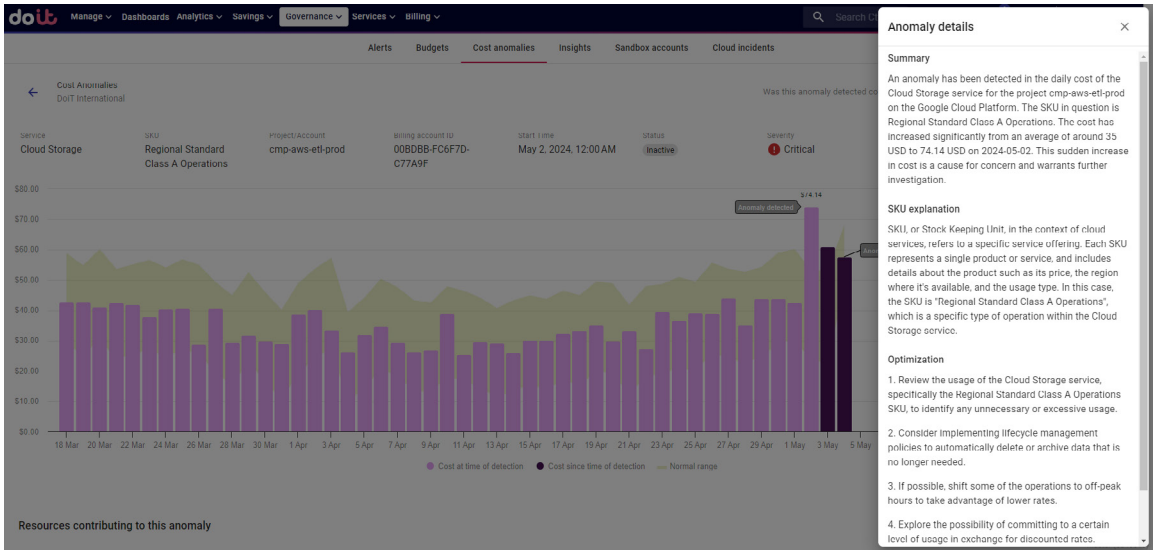
This is another area where DoiT Insights can help ease the burden on you and your cloud engineering / operations teams by continuously surfacing savings recommendations as they arise, thus alleviating the need for you to carve precious time out of your day to audit your AWS environment.

Additionally, you can leverage various GenAI tools within different solutions to further offload much of the mundane work that goes into ongoing cloud management. To do this within Cost Explorer, you can purchase access to Amazon Quicksight, which provides analysis of existing reports and BI tools to help visualize your cloud costs.

DoiT takes GenAI to the next level with Ava, which not only build reports and generate insights based on simple language queries, but is also able to provide in-depth analysis of any cost anomalies that get flagged through DoiT Anomaly Detection, thus lowering the time that it takes to investigate and remediate the root cause of any unplanned cost spikes.



Ava, DoiT Cloud Navigator's GenAI assistant



Ava, DoiT Cloud Navigator's GenAI assistant

Proper planning

As cliche as it may seem, the best way to optimize your AWS costs is by avoiding mistakes in the first place – namely, in the architectural design and development stages as you expand your cloud footprint.

Regardless of what stage you are in your cloud journey, as your business grows, odds are that you already have your eye on onboarding new services, providers, or technologies connected to or within your cloud environment. But not only are these expansion efforts a big burden on any engineering team on top of its regular operational efforts, but there's also a good chance that your team could benefit from additional expertise.

This is where Cloud Solve really delivers value, as the expansive team of global experts on staff have experience in everything from cloud migration and cost optimization, to more technically challenging expansion efforts like containers and Kubernetes or GenAI deployments. And in addition to dedicated consulting resources that can provide expertise and playbooks for your team to execute, DoiT also offers Accelerator programs, which are structured engagements designed to help you adopt new cloud services and build production-ready workloads faster than traditional timelines.

With that said, let's dive into some of the specific areas where your cost optimization efforts can have the biggest impact.

Compute Savings

Chances are that your biggest AWS cost driver is EC2, which can comprise anywhere from 50-80% of your overall cloud bill. This makes optimizing your compute spend the greatest opportunity to lower your overall public cloud costs.

Instance Types

One of the most fundamental ways to save on your compute workloads is make sure that you're using the most up-to-date instance types, which can often improve performance while still saving money on service costs. For example, AWS customers should consider Graviton-based processors for their EC2 workloads. These general purpose EC2 instance types deliver up to 40% better price performance over comparable x86-based instances. If you've already configured your workloads on previous generation instances, you can explore specific upgrade paths for the different machine families.

Given their flexibility, switching to Gravitron instances is applicable across many different AWS services such as RDS, OpenSearch, ElastiCache, Neptune, and more. However, to get ongoing monitoring for this kind of cost optimization throughout your environment, you will likely need to turn to a third-party solution like DoiT Cloud Navigator, which delivers continuous recommendations for cost savings as opportunities arise.

Scheduling & Rightsizing

One of the most common ways that compute costs can rise higher than necessary is by allowing your workloads to run constantly, even when not being utilized. This is especially damaging for applications that have usage fluctuations, such as a dropoff at night or over the weekend. Turning your compute workloads on and off based on when they're in demand by end users can ensure that you're not racking up unnecessary charges when no one is on your application. This can be done within the AWS console through Instance Scheduler or the Eventbridge scheduling service.

You can also save money by regularly rightsizing your workloads, which involves adjusting the resource allocation to match the actual workload requirements. The goal is to ensure that you're neither overprovisioning nor underprovisioning resources, which can have a significant impact on your monthly bill and overall cost efficiency. It's worth noting that rightsizing doesn't mean sacrificing performance. By selecting the right instance types and sizes that match your workloads' actual resource needs, you can often achieve better performance. Overprovisioned resources can lead to underutilization and increased costs.

To rightsize your AWS workloads effectively, you need to regularly monitor your resource utilization. You can use tools like AWS CloudWatch, Cost Explorer, and Trusted Advisor to gain insights into your resource usage patterns. Cost Explorer can help you identify instances that are overprovisioned, and you can then make informed decisions to resize or reconfigure them. Compute Optimizer is another service that provides rightsizing recommendations based on usage metrics, and supports multiple compute services such as EC2, ECS, and Lambda.

Increase your commitment discount coverage

While scheduling and rightsizing your compute workloads is useful, the most impactful way to save on compute costs is most likely through commitment-based discounts. AWS offers resource commitment discounts in the form of Savings Plans and Reserved Instances, both of which can be purchased in either 1- or 3-year terms, and which also come with some important differences in terms of buybacks and flexibility:

Characteristic	Standard RI	EC2 SP	Convertible RI	Compute SP
Change Availability Zone	✓	✓	✓	✓
Change Machine / Instance Size	✓	✓	✓	✓
Change Machine / Instance Family	✗	✗	✓	✓
Change OS	✗	✓	✓	✓
Change Region	✗	✗	✗	✓
Sellable on a Marketplace	✓	✗	✗	✗
Payment option	All upfront, partial, no upfront	All upfront, partial, no upfront	All upfront, partial, no upfront	All upfront, partial, no upfront

You can read more about the differences between these commitment types here, but the fact of the matter is that managing commitments and maximizing the savings available is a full-time job. One must take several different factors into account when forecasting compute usage – e.g. machine types, regions, cloud services, etc. – and then track usage and expiration dates to ensure that you're hitting the right milestones.

The best way to maximize your commitment coverage is to cover as much of your workloads as you can with 3-year RIs or Savings Plans, which offer 60-70% discounts as opposed to the 25-35% offered by 1-years. Bear in mind, of course, that 3-year commits are inherently riskier than 1-years simply because it's much harder to predict your workloads that far out. For companies that have the maturity and stability to do so, they could cover roughly half of their forecasted workloads with 3-year commitments, and then use 1-year commitments to round out their discount coverage.

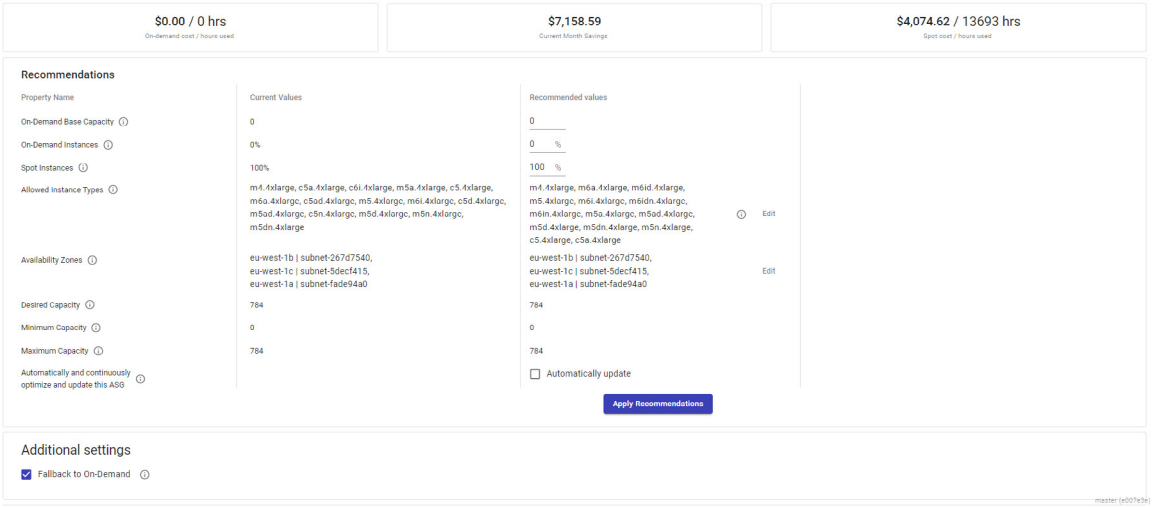
The beauty of a solution like DoiT Flexsave™ is that it automates the management of those1-year commitments to maximize the savings of any on-demand compute workloads (including EC2, Fargate, and Lambda) that aren't already discounted. Not only does this ease the FinOps management burden, but it also removes the risk of overcommitting to resources that you won't end up using.

Take advantage of Spot Instances

Like Savings Plans and RIs, Spot Instances can provide steep discounts on on-demand compute workloads, but with a heavy caveat – Spot workloads can be reclaimed by AWS with just a 2-minute warning. So while you can get even greater savings with Spot Instances with up to 90% discounts, they come with a great deal more risk, and should really only be used on fault tolerant instances like containerized workloads, stateless web servers, testing environments, or big data applications.

Spot Instances can be managed on AWS with Auto Scaling groups (ASGs), but deploying them naturally requires a certain degree of flexibility regarding the instance types and Availability Zones that you request. Why? Because none may be available to you at your target specifications. ASGs must also be manually configured and regularly tuned to ensure that your compute needs can be met consistently without meaningful interruptions. Apart from the manual, tedious nature of the process, if there's an incorrect configuration, it might not even work.

Given these risks and the effort involved, many practitioners choose to not even bother with Spot Instances. But DoiT Spot Scaling automates this process to remove the risk of interruptions and help to reliably run your workloads using Spot Instances. The tool automatically analyzes your ASGs to recommend best practice configurations, then replaces on-demand instances with the heavily discounted Spot Instances when applicable. And to eliminate the risk involved, Spot Scaling provides fallback to on-demand for situations where there isn't any spot capacity in the market.

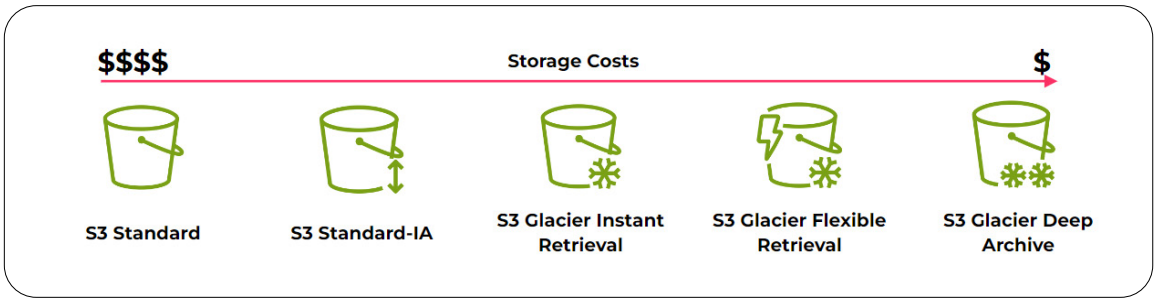


DoiT Spot Scaling

Storage savings

While not quite as dominant as compute on your monthly cloud bill, storage cost percentages can quickly ramp up as you scale, which makes it an area that should be regularly audited to ensure that costs are kept to a minimum.

AWS provides different classes of storage that vary in price depending on how frequently you need to access the data, which means that objects whose retrieval frequency has changed may still be stored in S3 Standard when they could be moved to Infrequent Access or Glacier.



To maximize the cost efficiency of your storage classes, you can use lifecycle policies to automatically transition data between different storage classes based on your access patterns. Make sure that you're also taking into account the amount of objects and their size, as retrieval and transfer prices are charged per GB. You can also use Amazon S3 Select to retrieve specific data from S3 objects, potentially reducing the amount of data transferred.

If you have unknown or unpredictable access patterns, then S3 Intelligent-Tiering is a great way to implement automation to save on your storage costs. While it's not a free service, the savings generated from it monitoring your access patterns and automatically moving objects infrequently-accessed objects to lower-cost access tiers can save you both time and money.

Using S3 VPC Endpoints can also help optimize data transfer costs and potentially reduce your AWS bill. VPC Endpoints allow you to securely access S3 storage buckets within your Virtual Private Cloud (VPC) without routing your traffic through the public internet, thus reducing or eliminating data transfer costs. This method is also deployable for the DynamoDB database service, which we'll explore later on.

Database and Data Warehouse savings

Similar to storage costs, data costs are one of the largest drivers of cloud spend, and can quickly pile up to create a dramatic impact on your monthly bill. AWS has several different data solutions available, and each comes with its own cost optimization strategies. Here we look at some of the most common cost drivers, and what you can do to mitigate them.

RDS

For many AWS customers, Relational Database Service (RDS) is their second-biggest cost driver after EC2, making it another prime opportunity for optimizations that will lead to significant reductions in your monthly bill.

A frequent culprit of RDS costs being higher than necessary is through the use of a Multi-Availability Zone (Multi-AZ) deployment, which is a high-availability configuration designed to enhance the resilience and availability of your database. In a Multi-AZ deployment, RDS automatically replicates your database in a primary AZ to a standby AZ, ensuring data redundancy and failover capability. The primary database is active and handles read and write operations, while the standby database can be quickly promoted to primary in the event of a failure.

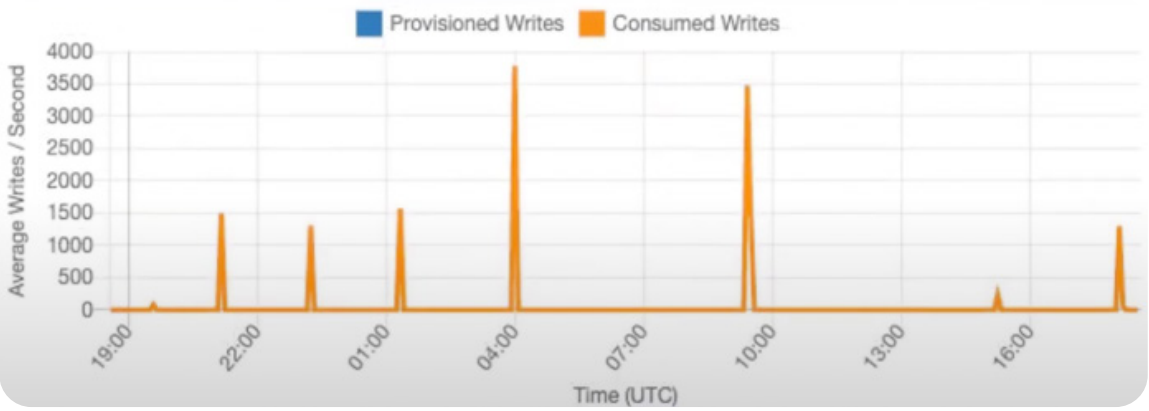
Yet while this is great for ensuring the reliability of your data, it can also get very expensive very quickly, so you should only use Multi-AZ for relevant use cases like production environments. Ultimately, the decision to use Multi-AZ deployments should be based on your application's availability requirements and budget considerations. While it may increase your AWS bill, the added reliability and reduced downtime can be a worthwhile investment for mission-critical applications.

Another way to save on your RDS costs is by using Graviton-based instance types whenever possible. Graviton is a family of EC2 instances that use custom-designed ARM processors, and are built to provide a cost-effective and energy-efficient alternative to traditional x86-based instances. This can lead to cost savings for your AWS infrastructure, making them a good choice for workloads where performance requirements can be met with ARM architecture. However, it's important to assess their suitability for your specific use case and verify compatibility with your software stack before migrating or deploying workloads on Graviton-based instances.

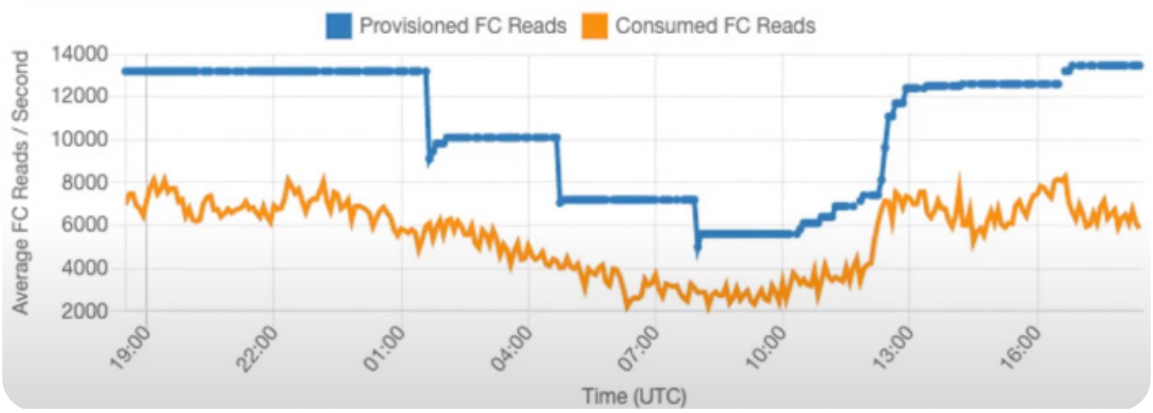
You can also leverage Aurora Serverless when applicable to further save on your RDS costs. Aurora Serverless is designed to simplify database management, reduce costs, and automatically adjust capacity based on your workload's demand. This can lead to significant cost savings, especially for workloads with variable usage patterns or when you want to eliminate the management overhead of manual database provisioning and scaling.

DynamoDB

If you're leveraging DynamoDB, AWS's managed NoSQL database service, you have a choice between the On-Demand or Provisioned modes for provisioning read and write capacity. With On-Demand mode, DynamoDB automatically handles capacity scaling, providing a serverless experience. You pay only for the actual read and write requests and the storage consumed by your table, but like with EC2 workloads, on-demand prices are higher and can quickly add up with consistent usage. This is why On-Demand mode is recommended if you have spiky usage patterns with traffic gaps, as seen here:



If, on the other hand, you have steady usage, then Provisioned mode allows you to manually specify the read and write capacity units for your DynamoDB tables. In this mode, you pay for the provisioned capacity, even if it's not fully utilized, as seen in this example below:



Provisioned capacity requires manual specification of capacity units and offers predictable costs. It's suitable for workloads with stable and predictable traffic patterns, but it may involve some capacity planning. You can also purchase reservations to further lower your throughput costs, with any excess throughput beyond your reserved capacity being billed at standard rates for provisioned throughput.

The choice between these capacity modes depends on your application's workload characteristics, traffic patterns, and budget constraints. You can also use a combination of both capacity modes for different tables within the same DynamoDB database to optimize costs and performance.

Amazon DynamoDB also provides a Time to Live (TTL) feature that allows you to automatically expire items from your tables after a specified period. Instead of manually deleting items, you can set a TTL attribute for each item, and DynamoDB will automatically remove items that have reached their expiration time. Not only does this save time (and therefore, operational costs) by removing the need for manual deletion, but it also saves on storage costs, as you won't be charged for storing expired items that are removed from the table.

Finally, if you have data that is infrequently accessed, or tables where storage is the dominant cost driver, you can consider implementing the Standard Infrequent Access (Standard-IA) table class. This can reduce your DynamoDB costs by up to 60%, and is ideal when you have long-term storage of data that is infrequently accessed, such as application logs, old social media posts, or ecommerce order history.

Redshift

AWS Redshift, a fully managed data warehouse, allows you to pause clusters based on usage patterns to optimize your AWS bill. While a Redshift cluster is running and actively processing queries, you are billed for the compute capacity (nodes) used during query execution; the more you use your cluster, the higher your query processing costs.

When you pause a Redshift cluster, it effectively shuts down the cluster, which means it's not actively processing queries, and you are not billed for its compute resources while it's paused. This is especially useful for development and test clusters, or clusters that are not continuously used.

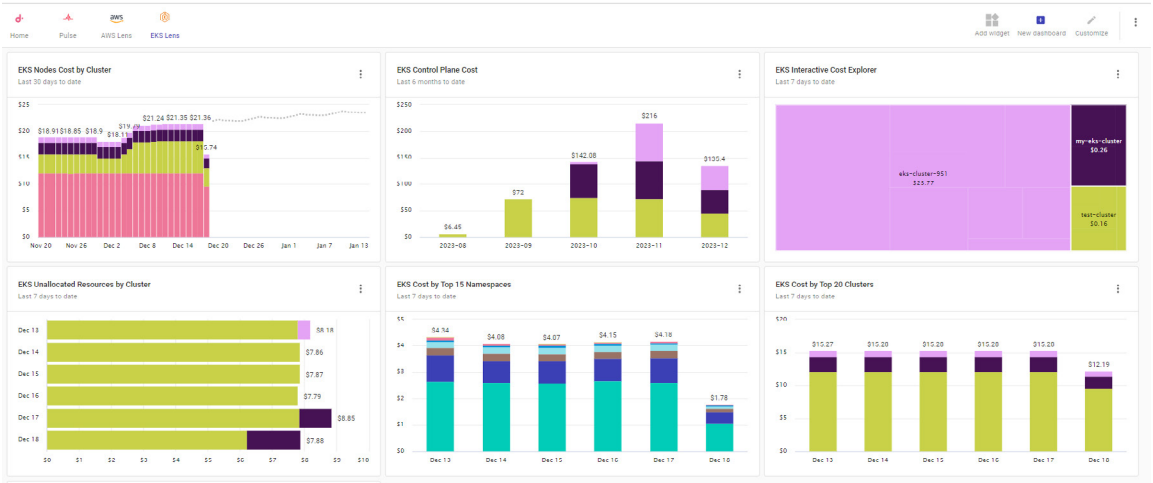
Data tiering is another effective cost optimization strategy wherein you separate your data into different storage tiers based on usage patterns, with frequently accessed data residing in high-performance storage and less frequently accessed data stored in more cost-effective storage. Redshift Spectrum allows you to implement data tiering and query data stored in Amazon S3 alongside your Redshift data warehouse. Since data tiering typically involves moving less frequently accessed data to lower-cost storage classes (e.g., S3 Glacier or S3 Deep Archive), it can result in cost savings compared to storing all data in high-performance storage.

Kubernetes savings

For companies that are mature enough in their cloud journey to have implemented containers and EKS orchestration, the ability to optimize them on an ongoing basis is crucial to getting the most value out of your investment.

And not only optimize them, but layer those costs with other business metrics to tie it back to the company's overall objectives.

DoiT provides this level of analytics and granularity through EKS Lens, an out-of-the-box dashboard that goes beyond what Cost Explorer can provide by integrating Kubernetes metrics with your AWS cost billing data. In doing so, it provides a unified, multi-cluster view for a comprehensive glimpse into your EKS usage and spend.

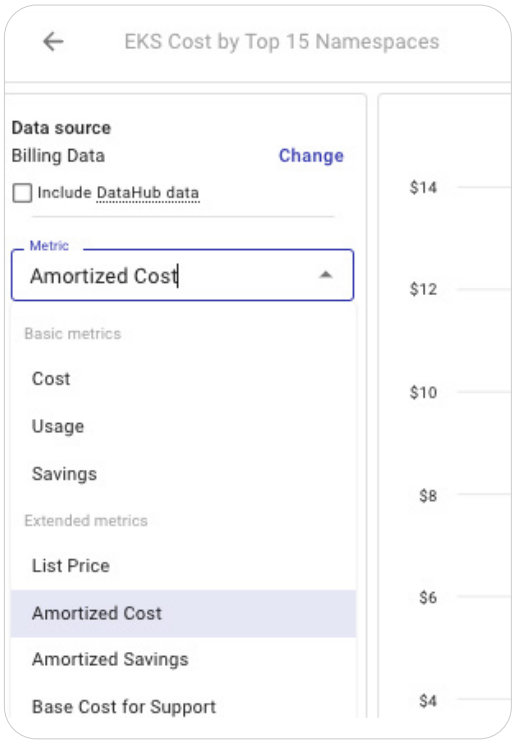


DoiT EKS Lens dashboard

It also enables you to view your costs as they amortize, providing even greater granularity of EKS clusters. It also makes it easier to accurately forecast your EKS clusters, thus enabling more reliable decisions about your budgeting and savings strategies.

Regardless of the tools you use, when trying to maximize the return on your Kubernetes investment, it's helpful to focus on three key drivers of excess EKS costs:

- Auto Scaling
- Rightsizing
- Spot Instances



Auto Scaling

Implementing auto-scaling in EKS can help optimize your AWS bill by dynamically adjusting the number of nodes in your cluster based on workload demand. Auto-scaling ensures that you have enough resources to handle your application's load efficiently without over-provisioning. To maximize your cost efficiency, it helps to regularly review your application's performance, monitor auto-scaling actions, and fine-tune configurations to achieve the right balance between performance and cost.

This includes downscaling to reduce the number of replicas for your EKS pods. When demand for your application decreases, the auto-scaler might initiate a downscaling action, resulting in a reduction in the number of running pod replicas.

Rightsizing

You can start rightsizing your EKS workloads by monitoring your resource usage to understand the utilization patterns. You can then optimize pod resources by allocating the appropriate CPU and memory resources. And remember that resource requirements can change over time; implementing rightsizing in your EKS workloads is an iterative process that requires continuous monitoring and adjustment, so regularly review your application's performance and adjust your rightsizing strategies accordingly.

Spot Instances

Consider using Spot Instances for nodes in your EKS cluster, especially for workloads with flexible resource requirements. Spot Instances can provide significant cost savings compared to On-Demand Instances, but as previously discussed, they come with the trade-off of potential interruptions. Here again is where a solution like DoiT Spot Scaling can greatly increase the efficiency of your Spot usage by automating the process to maximize your savings, and falling back to on-demand workloads when there is no Spot inventory available.

Savings through Cloud Governance

For companies that are mature enough in their cloud journey to have implemented containers and EKS orchestration, the ability to optimize them on an ongoing basis is crucial to getting the most value out of your investment.

While all of the services we've already discussed should be part of your Well-Architected reviews and regular cost optimization audits, the day-to-day monitoring of cost spikes is also a key part of any FinOps practice. As your AWS investment grows and your environment becomes more complicated and widespread, the potential for significant cost anomalies increases. Not only can these anomalies have a significant negative impact on your business, but they can also be just the tip of the iceberg that indicates a much larger problem within your AWS environment.

Cloud cost anomalies are something that need to be monitored for and alerted on regardless of how big or small your company and cloud environments are. The last thing you want to do is think that you've got your costs under control, only to be faced with a far-bigger-than-expected bill at the end of the month.

But even greater than the immediate impact on your cloud bill, cost anomalies can also be the first symptoms that can reveal problems in the underlying infrastructure, making it all the more important to investigate and remediate them as soon as possible to avoid even larger problems down the line.

Infrastructure

- [ ] Are your workloads misconfigured?
- [ ] Do your instances need to be rightsized?
- [ ] Has someone on your team improperly provisioned resources?

Software

- [ ] Is there a bug in your code that's affecting your cloud usage?
- [ ] Was there a software update that threw off your cloud configurations?

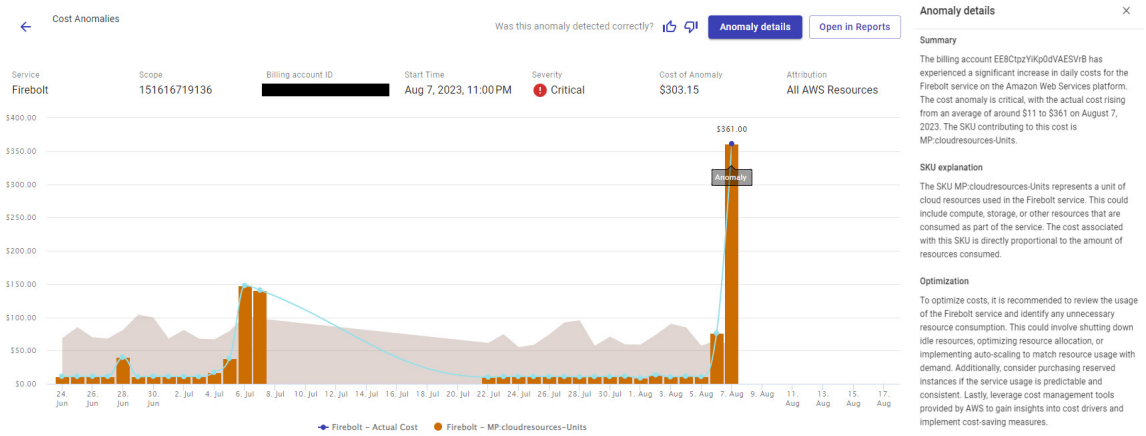
Security

- [ ] Has your cloud account been compromised?

Of these three potential areas of concern, a security issue is probably the most important to get under control because of the business risk that compromised credentials can introduce to your entire tech stack, along with wider financial and legal ramifications. Yet one of these three areas can have the potential to spiral out of control if not dealt with in a timely manner, which only increases the importance of early detection and mitigation.

DoiT helps mitigate these with intelligent Anomaly Detection software that uses a combination of statistical metrics and time series modeling to filter samples for relevance, compute their excess, and determine the severity of anomalies. The system is tuned to optimize alert effectiveness by ensuring all anomalies are identified, but only notifying customers of significant cases, thus avoiding alert fatigue.

Once an alert is sent, it works hand-in-hand with DoiT Cloud Analytics, providing reports combined with actionable insights powered by Ava (DoiT's GenAI solution within Cloud Navigator) that customers can use to quickly investigate the problem at its source.



Cost Optimization with DoiT

While the methods covered in this book may seem daunting, especially on an ongoing basis, there are even more cost optimization opportunities that could be available depending on your specific AWS environment.

Unearthing those opportunities not only requires extensive cloud expertise and vigilant monitoring, but also the necessary tools to automate and expedite the process whenever possible, and which may not be available in the native AWS tooling.

The previously mentioned Well-Architected review that the DoiT services team regularly conducts is just one of the many outputs available to DoiT customers. These also include on-demand training and support, marketplace acceleration, playbooks for implementing new services such as Kubernetes, and FinOps readiness assessments.

In addition to those services, DoiT customers can access the DoiT technology portfolio, which includes cost allocation and Cloud Analytics solutions, cloud governance through intelligent Anomaly Detection, Budgets, and Alerts, and automated commitment management through DoiT Flexsave.

To learn more about the different ways to leverage the DoiT technology and services offerings, get in touch with one of our experts.

Chat with us

Meet doit

DoiT works alongside cloud driven organizations to optimize your cloud use so you can focus on business growth and innovation.

We simplify your most important cloud challenges with the tools and expertise to help you buy, manage, and optimize cloud usage and costs. DoiT delivers multicloud procurement advantage, world-class expertise to solve complex challenges, and full-service FinOps solutions to navigate spend.

Accelerate cloud optimization

Control spend & performance

Enhance skills & capabilities

Learn more at DoiT.com.