nOps™

EBOOK

The Definitive Guide to AWS Rightsizing

www.nops.io

Table of Contents

Introduction

Controlling cloud costs has overtaken security as the top cloud challenge organizations face, with 82% of organizations currently highlighting it as their top concern amid rising bills and economic uncertainty. [1]

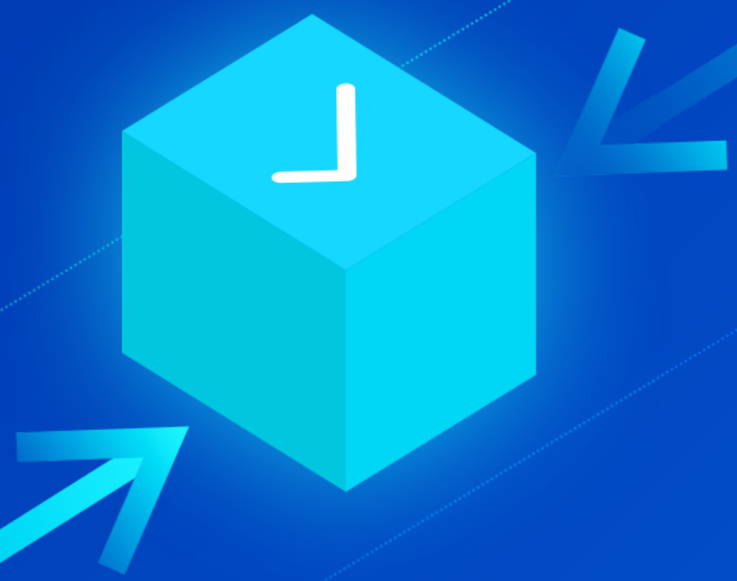
At nOps, we help customers optimize over $1.5 billion in AWS spend. We have observed that many companies are currently over-provisioning their cloud resources, presenting significant opportunities for cost reduction.

While the cloud's ability to quickly provision resources is a key benefit, it can also lead to unnecessary spending when usage decreases — especially when it comes to EC2 resources. That’s where rightsizing comes in. By analyzing historical usage and performance, you can identify and rightsize instances that are not consuming all the resources that are currently available to them.

Rightsizing these resources to their proper level goes a long way toward taming cloud costs while ensuring performance and stability. In this ebook, we’ll take you through a comprehensive, hands-on guide with step-by-step instructions, effective strategies, and best practices for rightsizing.

[1] “*Flexera 2023 State of the Cloud: Report.” Flexera 2023 State of the Cloud | Report, info.flexera.com/CM-REPORT-State-of-the-Cloud. Accessed 30 Jan. 2024.*

Introduction to Rightsizing



Introduction to Rightsizing

Rightsizing, sometimes spelled Right Sizing, is the process of matching instance types and sizes to your workload performance and capacity requirements at the lowest possible cost.[2] It also includes reviewing deployed instances and identifying opportunities to eliminate or downsize without compromising capacity or other requirements, resulting in cost savings.

01. The Importance of Rightsizing

Rightsizing is one of the most effective ways to control cloud costs. It involves analyzing instance performance, usage needs, and patterns to establish a usage baseline — then turning off idle instances and re-sizing instances that are overprovisioned or poorly matched to the workload.

Even if you rightsize workloads initially, performance and capacity requirements for workload change over time. If you are not constantly rightsizing workloads, you are operating AWS like a traditional data center — it not only leads to unnecessary costs but also hinders optimal performance.

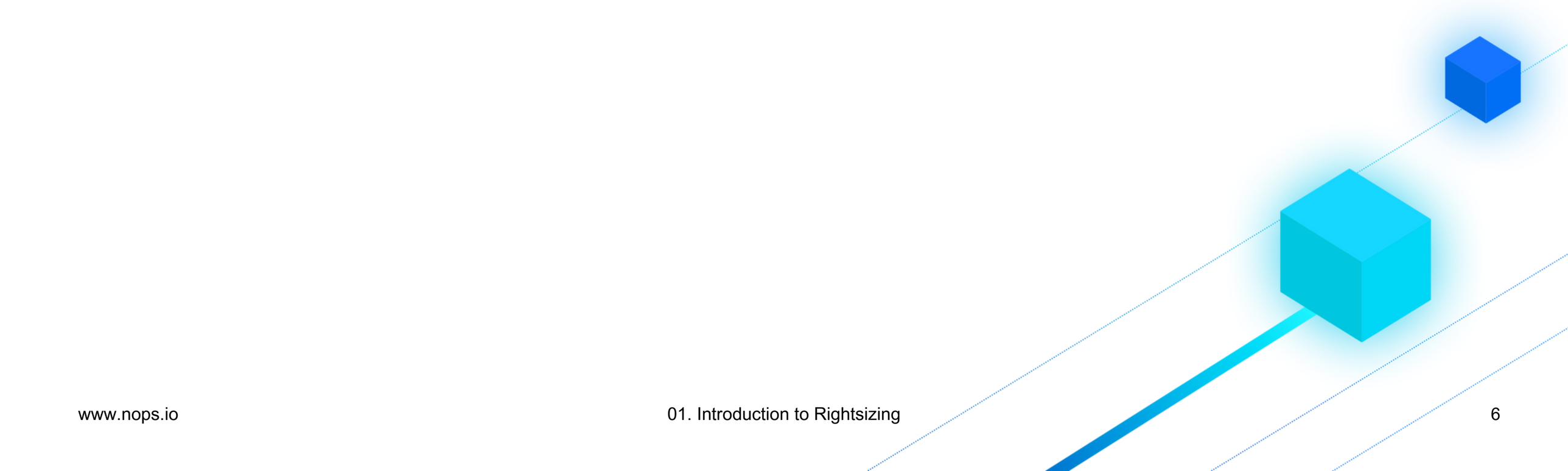
[2] Wilkins, M. (2021). AWS well-architected framework: Cost optimization pillar. Amazon. https://aws.amazon.com/aws-cost-management/aws-cost-optimization/right-sizing/

02. The Challenges of Rightsizing

One of the challenges with rightsizing is that developers are afraid that it might break their workloads. That’s why it’s important to have accurate recommendations that allow you to optimize for both operational efficiency and cost.

Vast amounts of rightsizing recommendations are available through tools such as AWS Trusted Advisor. Yet, the problem is that rightsizing recommendations are frequently wrong. Reliable recommendations require granular historical data on memory, utilization, network bandwidth, volume size, and more. This data is difficult to collect and analyze, and there isn't a single source for all of the metrics that you need. As a result, rightsizing recommendations are often untrustworthy — and engineers don't act on them as a result.
The other major challenge is that rightsizing is a continual and time-consuming process. Cumbersome resource investigation, ticketing processes, and manually editing Terraform means that a right sizing initiative can be an IT resource time sink.

Accurate recommendations only take you so far; you have to make it easy for developers to take action. In this ebook, we'll cover practical information with how-to guides, screenshots and best practices for rightsizing effectively.



Key Metrics for Rightsizing

Key Metrics for Rightsizing

To effectively rightsize your AWS resources, you'll need to analyze specific metrics.

Most companies use Datadog or CloudWatch for monitoring; as such, we've included the relevant key metrics you'll need for rightsizing. If you are using a different monitoring tool, you can look for similar metrics.

Metric Category	How It's Used	CloudWatch Metric(s)	Datadog Metric(s)
CPU Utilization	Evaluate the CPU usage to determine if a smaller instance could suffice.	**CPUUtilization**: Percentage of CPU utilization.	We recommend you use the CloudWatch metric — the closest Datadog metric to use is **cpu.idle**, which approximates CPU utilization when inverted.
Memory Utilization	Memory metrics are crucial for accurate rightsizing, especially for certain instance types. Tools like AWS CloudWatch and third-party solutions like Datadog are invaluable.	**mem_used_percent** (custom metric): Percentage of used memory.	**mem.used**: Amount of RAM currently in use.
Network Utilization	Assess disk I/O and network throughput to ensure your instances match your actual usage patterns.	**NetworkIn**: Total bytes received on the network. **NetworkOut**: Total bytes sent on the network.	**net.bytes_rcvd**: Total bytes received over the network. **net.bytes_sent**: Total bytes sent over the network.
Storage Utilization	Focus on ephemeral disk usage to gauge whether the current allocation aligns with your actual needs.	**disk_used_percent** (custom metric): Percentage of used disk space.	**disk.used**: Amount of disk space currently in use.

Please note that mem_used_percent and disk_used_percent are only available through CloudWatch Agent (they are not included in the free version of CloudWatch).

01. Establishing Baselines and Using Metrics

The first step in rightsizing is to monitor and analyze your current use of services to gain insight into instance performance and usage patterns. At nOps, we recommend using at least a 15-day period (or ideally, 30 days) to establish a reliable baseline for rightsizing recommendations. Let’s discuss a few key guidelines:

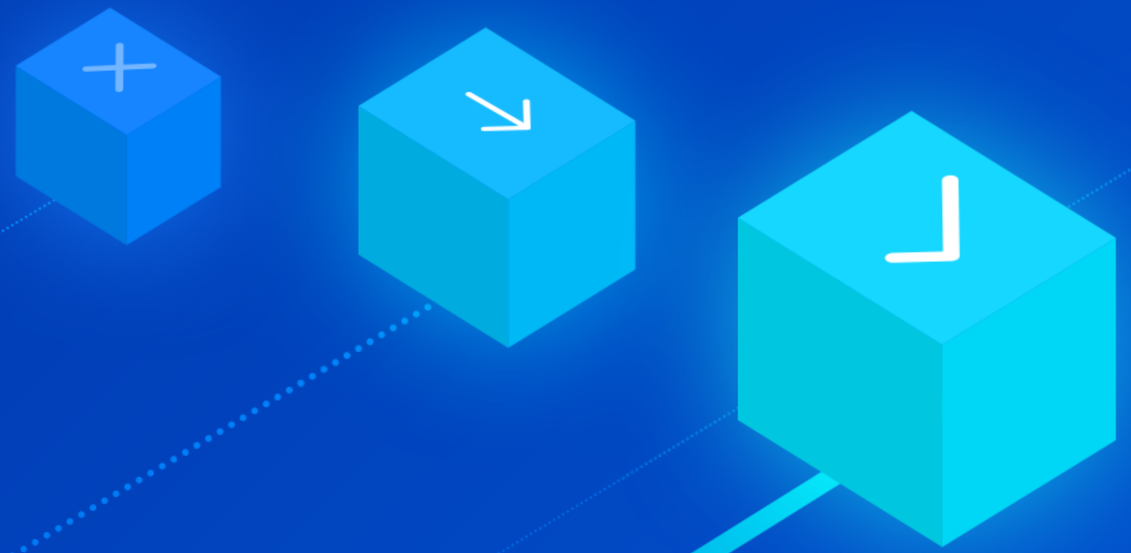
- For the most accurate recommendations, increase the granularity of your metrics and observing over a longer time period (particularly if your workload is variable).

- We’ll focus on maximum utilization rather than average or mean/median utilization. This is a slightly more conservative, but safer path for production workloads. For UAT or other non-critical workloads, you may wish to be more aggressive.

- Your instances may not be constantly running. If this is the case, the minimum period of data collection would be to observe the equivalent of 15 days (i.e. 360 hours) within a 30-day period.

- The general “rightsizing rule” is to compare your usage against two baselines: your current instance type and a smaller instance type. If your usage is consistently less than 80% of the smaller instance's baseline, consider downsizing. We’ll delve into the details of how this is done for each metric later in the ebook.

How to Rightsize: A Step-By-Step Guide



How to Rightsize: A Step-By-Step Guide

We'll take you through the full process of rightsizing for both CloudWatch and Datadog. If you are using a different monitoring solution, you can adapt these steps accordingly.

01. Rightsizing with CloudWatch

Amazon CloudWatch gives us basic metrics by default which are CPU, network and storage. However, reliable recommendations also need to consider memory. AWS offers a paid service (or "agent") that needs to be installed to get that fourth metric, called "CloudWatch Agent"

i. Integrate CloudWatch Agent.

AWS has an easy-to-follow guide including commands to run on your EC2 instance.

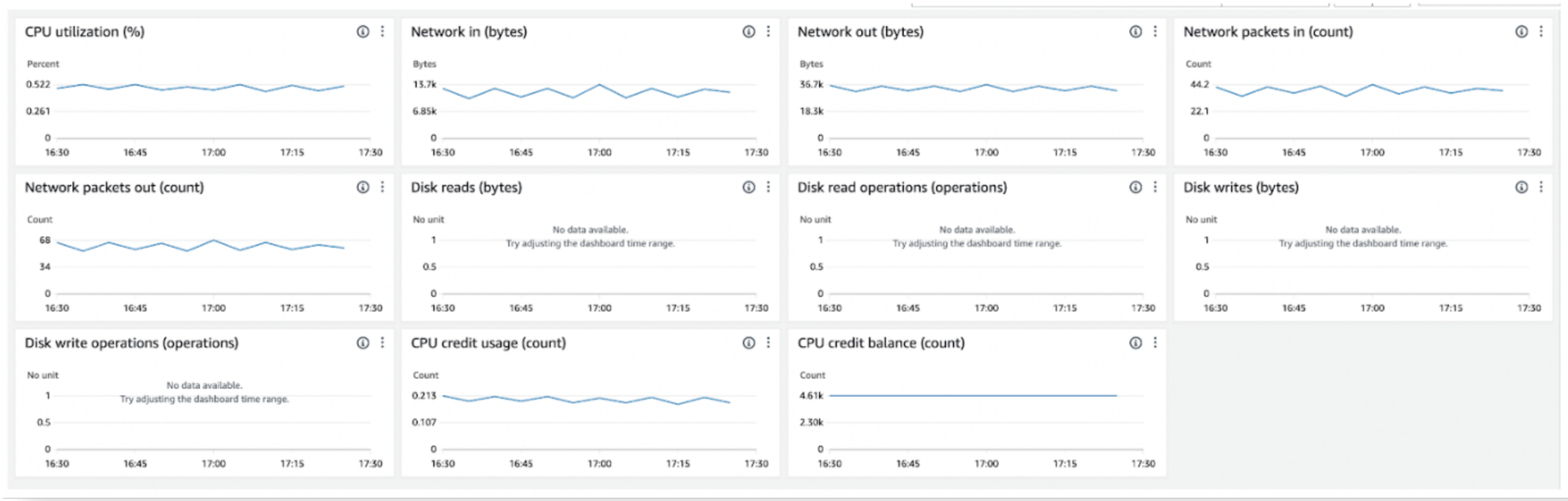
Command line for installing CloudWatch

ii. Wait 15+ days.

Once you've successfully configured the AWS CloudWatch agent, it will now start giving us memory utilization data. As stated, you need to collect the memory metrics for a minimum of 15 full days (or for 360 hours in a 30-day period) to be able to make statistically significant recommendations.

iii. Review metrics in Cloudwatch:

Once the metrics are collected, they can be reviewed in the CloudWatch dashboard. Evaluate 4 key metrics: vCPU, Memory, Storage and Network. These metrics are tagged as custom metrics. CloudWatch recommendations support multiple Operating Systems, including Windows and Linux, working at an OS level to bring detailed metrics.



CloudWatch metrics used for rightsizing

iv. Perform rightsizing calculations.

Compare the current usage of each metric against two baselines: one for your current instance type and another for the next smaller instance type. If your maximum usage is less than 40% of the current type, or up to 80% of the smaller instance's baseline (i.e., you're using 20% less than what the smaller instance typically supports), it's safe to consider downsizing to that smaller instance type.

Here are the metrics and formulas used to rightsize.

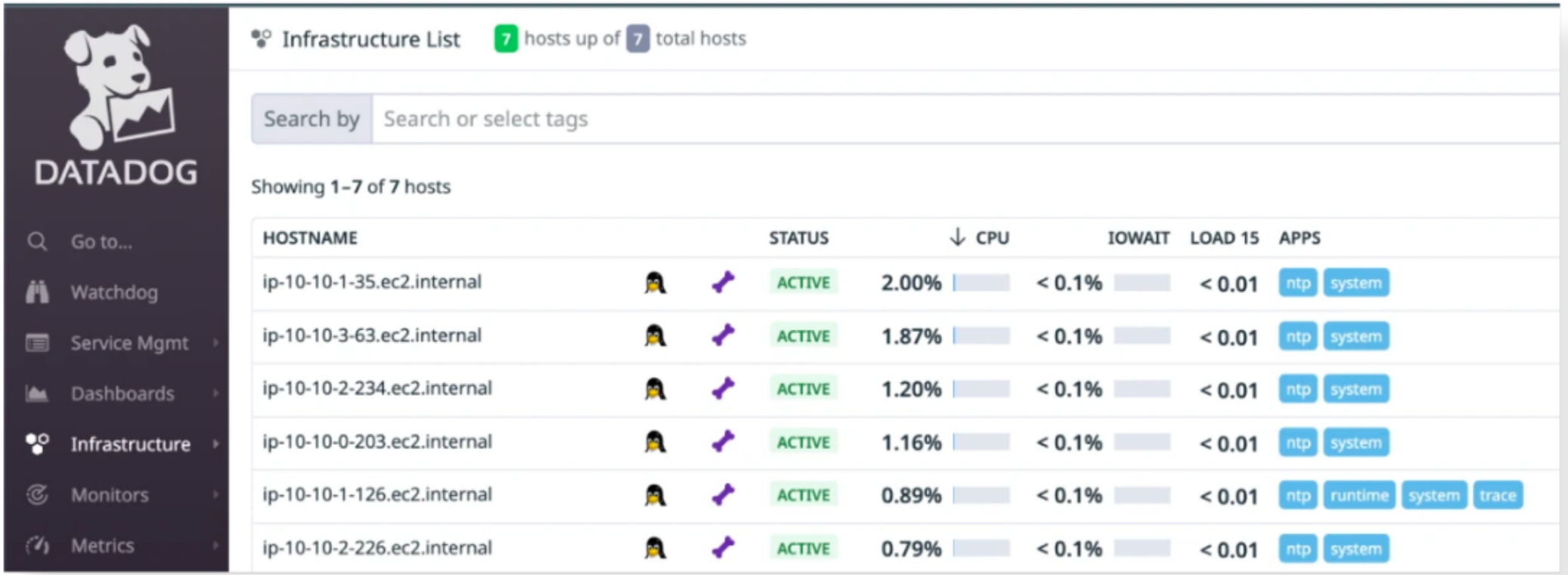
Metric	What it represents	Period to use	Metric type	Formula	Notes
CPUUtilization	CPU Utilization percentage.	Last 30 days	Max	Rightsize if CPUutilization is at or below 40%	
Mem_used_percent	Memory being used	Last 30 days	Max	Rightsize if mem_used_percent is below 80% of the lower instance type	
NetworkIn **NetworkOut**	Network utilization (important for workloads relying on high bandwidth)	Last 30 days	Max	Rightsize if NetworkIn + NetworkOut is below 80% of the lower instance type	This depends on instance type — ensure that you put everything in the same scale, whether bytes per second, bits per second, or percentage
DiskReadOps **DiskWriteOps**	Utilization of ephemeral (nonpersistent) storage — not to be confused with EBS (persistent) storage	Last 30 days	Max	Rightsize if DiskReadOps + DiscWriteOps is below 40%	

02. Rightsizing with Datadog

i. Sign up for Datadog.

ii. Integrate your cloud resources

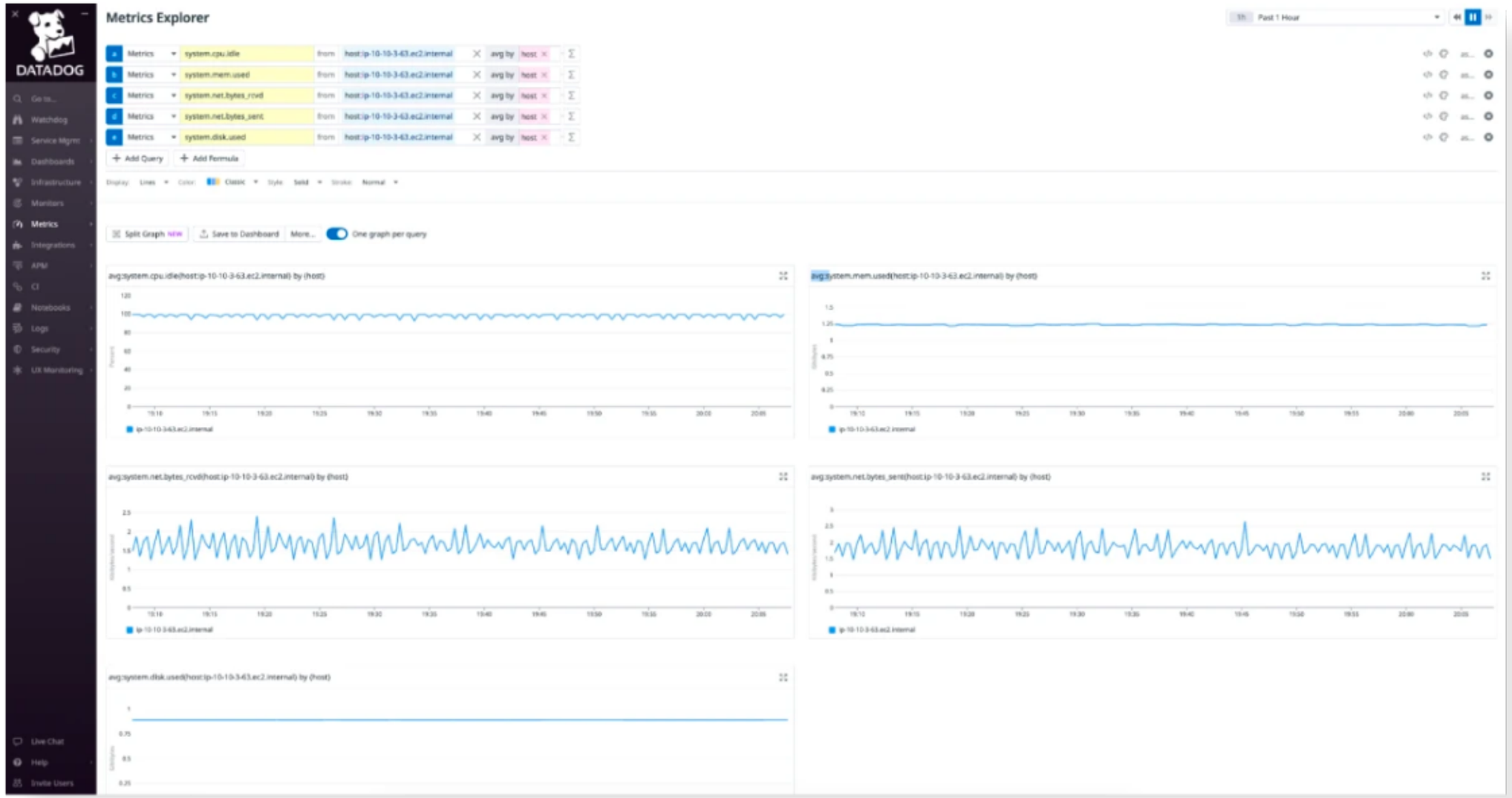
(e.g. EC2, RDS) with your Datadog account. Within 15-30 minutes, data will start flowing into Datadog



iii. Wait 15+ days.

As stated above, after integration, we generally recommend that you wait at least 15-30 days to get the data you need for solid recommendations.

iv. Go to the **Datadog dashboard → Infrastructure → HOSTNAME → Metrics**



v. Evaluate 4 key metrics: vCPU, Memory, Storage and Network.

Compare the current usage of each metric against two baselines: one for your current instance type and another for the next smaller instance type. If your maximum usage is less than 40% of the current type or up to 80% of the smaller instance's baseline (i.e., you're using 20% less than what the smaller instance typically supports), it's safe to consider downsizing to that smaller instance type.

Visible in the screenshot above, here are the Datadog metrics that you need to pay attention to in order to make good rightsizing decisions.

Metric	What it represents	Period to use	Metric type	Formula	Notes
CloudWatch CPUUtilization or Datadog cpu.idle	CPU Utilization percentage	Last 30 days	Max	Rightsize if CPUtilization is at or below 40% (or 1/cpu.idle is below 40%).	We recommend you use **CPUUtilization** which is a free CloudWatch metric, rather than Datadog. If you prefer, you can use **cpu.idle**, which when inverted approximates CPUUtilization.
mem.used	Memory being used	Last 30 days	Max	Rightsize if mem.used is below 80% of the lower instance type	
net.bytes_rcvd net-bytes_sent	Network utilization (important for high bandwidth workloads)	Last 30 days	Max	Rightsize if NetworkIn + NetworkOut is below 80% of the lower instance type	This depends on instance type — ensure that you put everything in the same scale, whether bytes per second, bits per second, or percentage.
disc.used	Utilization of ephemeral (nonpersistent) storage — not to be confused with EBS (persistent) storage	Last 30 days	Max	Rightsize if DiskReadOps + DiscWriteOps is below 40%	

Note that the above recommendations hold as long as you are planning to rightsize within the same instance family.

Additional Best Practices for Rightsizing

Additional Best Practices for Rightsizing

Here are more tips to keep in mind as you rightsize your EC2 instances.

01. Follow AWS recommendations for CPU and memory

AWS's general rule for EC2 instances is that if your maximum CPU and memory usage is less than 40% over a four-week period, you can safely reduce capacity.

For compute-optimized instances, some best practices are to:

- Focus on very recent instance data (such as the previous month), as old data may not be actionable
- Focus on instances that have run for at least half the time you're observing
- Ignore burstable (T2) instance families, as they are designed to run at a low CPU % for extended periods

02. Select the right instance type and family

AWS offers hundreds of instance types at different prices, comprising varying combinations of CPU, memory, storage, and networking capacity. You'll want to select for the lowest price per unit of the metric most important for your workload.

You can rightsize an instance by migrating to a different machine within the same instance family or by migrating to another instance family.

When rightsizing outside the instance family, you also need to consider CPU architecture, storage type, storage speed, and other factors.

- Virtualization type: The instances must have the same Linux AMI virtualization type (PV AMI versus HVM) and platform (EC2-Classic versus EC2-VPC).

- Network: Instances unsupported in EC2-Classic must be launched in a virtual private cloud (VPC).
- Platform: If the current instance type supports 32-bit AMIs, make sure to select a new instance type that also supports 32-bit AMIs (not all EC2 instance types do)

03. Turn off idle instances

One easy way to reduce operational costs is to turn off instances that are no longer in use. The AWS guideline states that it's safe to stop or terminate instances that have been idle for more than two weeks.

Some key considerations for terminating instances include (1) who owns it (2) what is the potential impact of terminating (3) how hard is it to recreate.

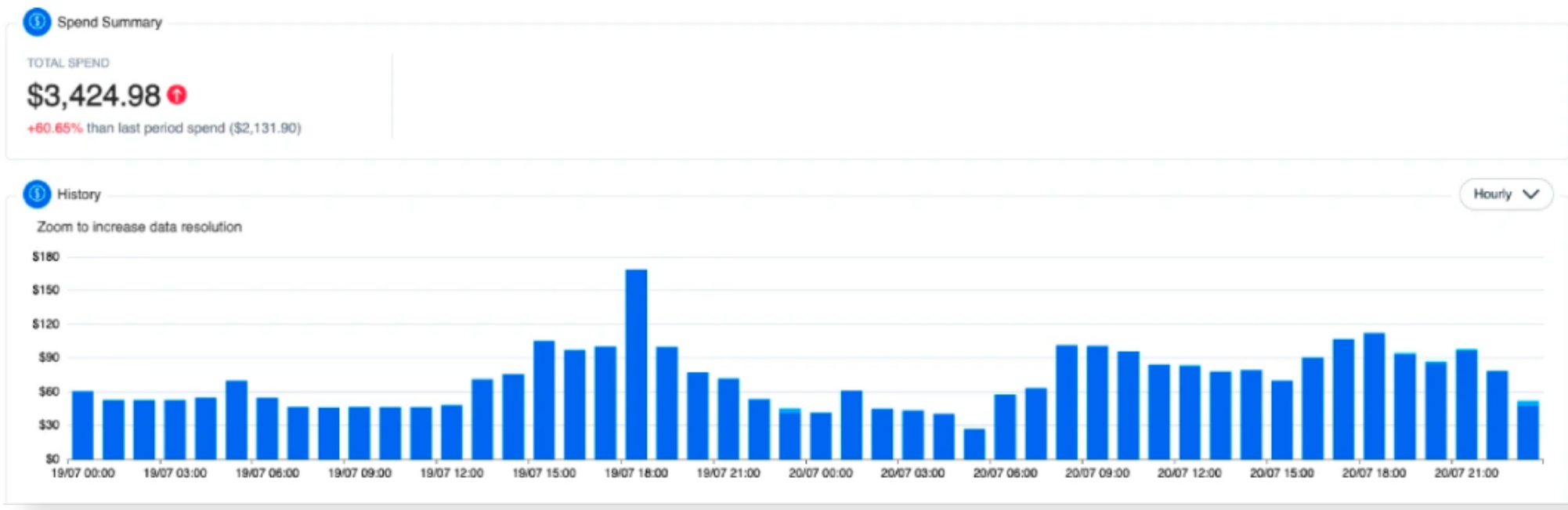
Another simple way to reduce costs is to stop instances used in development and production during hours when these instances are not in use. Assuming a 50-hour work week, you can save 70% by automatically stopping dev/test/production instances during non-business hours.

Many tools are available for scheduling, including Amazon EC2 Scheduler, AWS Lambda, and AWS Data Pipeline. Or, third-party tools such as nOps use AI to learn your usage patterns and automate the process for you.

04. Monitor your resource usage over time

To achieve cost optimization, rightsizing becomes an ongoing process. Even if you rightsize workloads initially, changing performance and capacity requirements can result in underused or idle resources that drive unnecessary AWS costs.

As you monitor current performance, identify the following usage needs and patterns so that you can take advantage of potential rightsizing options:



Example workload, with hourly granularity

- **Steady state** – When workloads maintain consistent levels over time, and you can forecast compute needs accurately, Reserved Instances offer significant savings.
- **Variable, but predictable** – If your load changes predictably, consider EC2 Auto Scaling to handle predictable fluctuations in traffic.
- **Dev/test/production** – These can generally be turned off during non-business hours. (You'll need to rely on tagging to identify dev/test/production instances.)
- **Temporary** – For temporary workloads that have flexible start times and can be interrupted, consider using an Amazon EC2 Spot Instance.

Continuous Rightsizing

Continuous Rightsizing

nOps integrates with the two industry-leading monitoring solutions, AWS CloudWatch and Datadog, for effortless rightsizing savings. We automatically analyze every EC2 instance in your environment for recommendations, which you can apply with one click on the platform.

Real-time coverage of resource-level insights such as memory, CPU, network bandwidth and storage are fed through nOps's state-of-the-art ML engine for the **best rightsizing recommendations available on the market**.

Rightsize with nOps for:

The most trustworthy rightsizing recommendations.

Because nOps automatically collects and analyzes highly granular data, recommendations are 100% accurate and reliable — so engineers can act on them with the utmost confidence that workloads won't be disturbed.

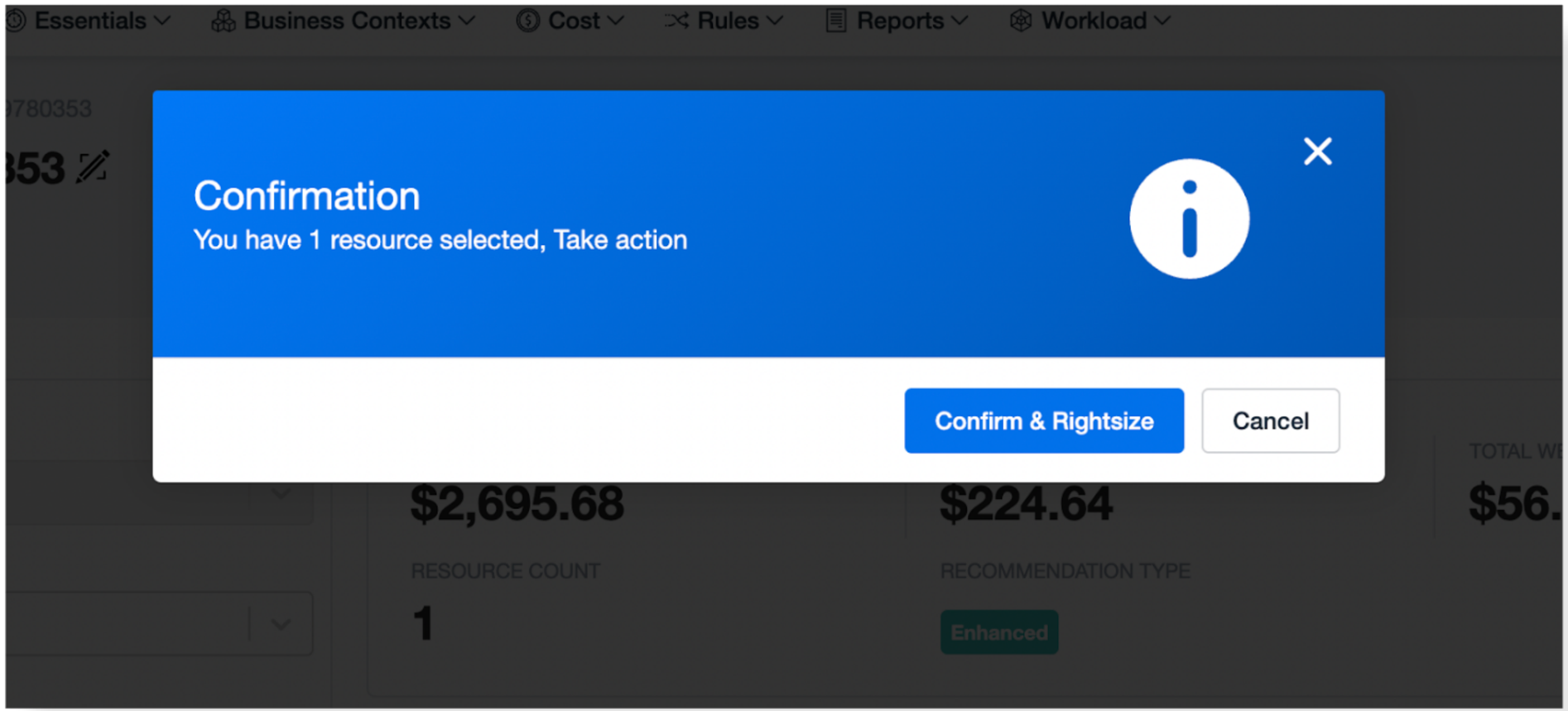
Significant time savings.

nOps integrates with EventBridge, automating the complex and time-consuming rightsizing process into a single click — freeing up engineers to build and innovate.

Up to 50% in immediate cost savings.

When engineers don't act on rightsizing recommendations, underutilized and idle resources continue to drive unnecessary AWS costs. nOps make it completely pain-free, safe and effortless for engineers to actually act on recommendations and start saving.

One-click rightsizing, right from nOps using EventBridge



About nOps

About nOps

nOps is a certified AWS Select Tier Services Partner and AWS Marketplace Seller. nOps helps companies automatically optimize any compute-based workload. Our mission is to make it faster and easier for engineers to take action on cloud cost optimization, so they can focus on building and innovation.

nOps processes over $1.5 billion in cloud spend and was recently ranked #1 in G2's cloud cost management category. Join our customers using nOps to cut cloud costs and leverage automation with complete confidence by **booking a demo** today!