pepperdata

SOLUTION BRIEF

Reclaim Waste and Reduce Cloud Costs With Autonomous Big Data Optimization

Most solutions for minimizing waste in your big data clusters leave money on the table. They're a time sink, too. Pepperdata Capacity Optimizer is a radical new solution for minimizing waste and reducing cloud costs autonomously. Free your developers from the manual busy work of optimization so they can focus on innovation.

Enterprises are aggressively moving their workloads to the cloud. Gartner estimates that by 2025, over half of IT spending in key segments will shift to the public cloud.

At the same time, organizations are struggling to handle growing cloud spending. They are over budget for cloud spend by an average of 23 percent, yet they expect cloud spend to increase by 47 percent over the next year.

Faced with these budget overruns, enterprises are adopting a variety of approaches to control costs, including minimizing waste, reducing consumption, and shifting to less expensive services to eliminate the estimated 30 percent of cloud computing services that go wasted.

The cloud's appeal in providing on-demand, elastic, and highly scalable computing resources contributes to this waste. The cloud also tends to make it challenging for developers to identify and eliminate waste—until now.

Systems are Underutilized

Many production data systems are subject to fluctuating parameters, such as cluster configurations, instance types, varying data size, skew, volume of jobs, and a constantly changing codebase. Larger enterprise deployments usually have thousands of jobs running on multi-tenant containers. Engineers tend to overprovision resources in an attempt to create buffer for unexpected complexity. This overprovisioning invariably results in underutilization and therefore waste.

Almost Every Factor That Makes the Cloud Appealing Contributes to Cost Overruns

- Usage-based billing
- Elasticity
- High scalability
- Dynamic resource changes
- New financial models
- On-demand computing

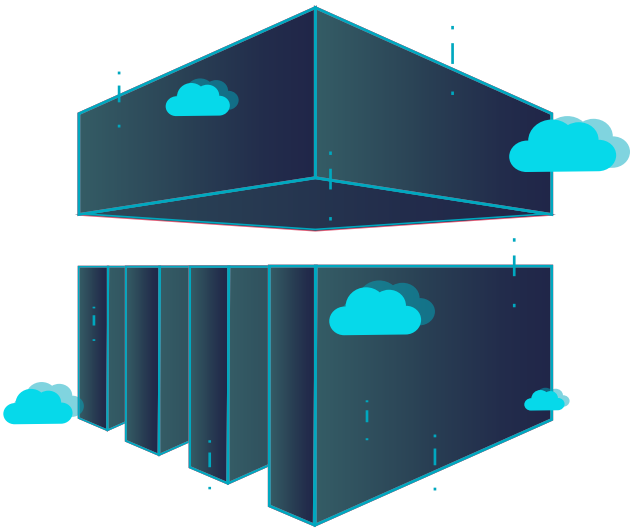
Even The Most Dedicated Engineer Can't Keep Up

Once applications are up and running, cost and resource problems begin to multiply. The load and available resources in a cloud environment change dynamically in real time as jobs execute and terminate and new ones are run. The volume, variety, and velocity of data flowing through the modern data stack is such that no developer can keep on top of all of it. No matter how deep the visibility into runtime resource utilization, the complexity and speed of the modern data stack defies manual tuning. Consequently, whatever cannot be manually tuned—which could be a sizable portion of a cluster's resources—results in waste.

- **Large datasets**
- **Intense computing**
- **Autoscaling**
- **Batch jobs**

Most Cloud Autoscaling is Massively Inefficient

Cloud autoscalers allow a cluster to automatically tune the number of instances based on the needs of the job. It evaluates cluster metrics to determine if more or less resources are needed. The general message about autoscaling is that it optimizes the cluster for your workload and minimizes costs.



However, in reality, default autoscalers are highly inefficient and can often contribute to significant waste. This is because autoscaling is not a "set and forget" solution—it needs constant tuning to decide optimal settings for the application. Our studies have shown that default autoscalers can increase costs by as much as 30 percent.

Imagine a cluster with two nodes. Let's say an application requests two nodes. The application may end up using only eight cores, but the cluster autoscaler does not know that. As more applications are submitted requesting more cores, the autoscaler will add more instances even though the existing instance is only 50 percent utilized. If new applications also use just a fraction of the allocations, the new instances will be also underutilized. The result is many more wasteful instances and, ultimately, inflated cloud bills.

Pepperdata Capacity Optimizer:
Autonomous Optimization Without Manual Intervention

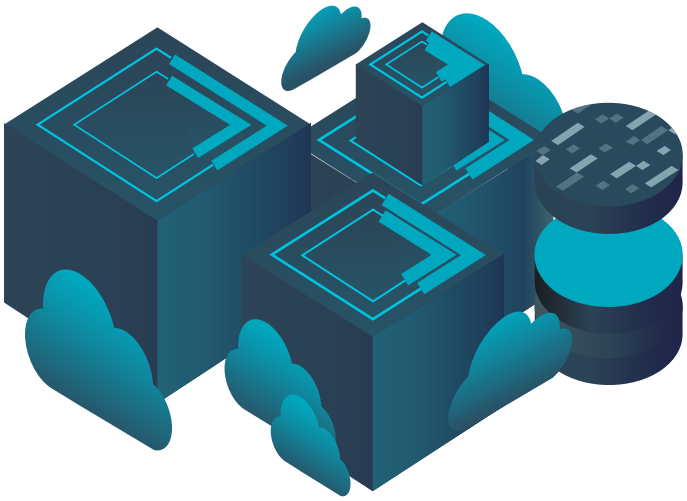
Capacity Optimizer from Pepperdata is a radical new way to minimize the inherent waste of big data and Kubernetes systems in the cloud and increase utilization in on-prem environments. It is the only solution that operates autonomously, continuously, and in real time to solve the price/performance problems of big data in both YARN- and Kubernetes-based environments.

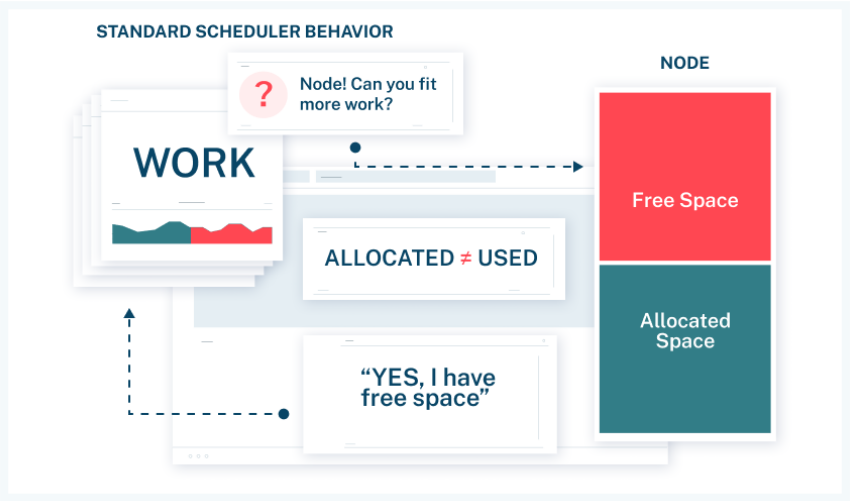
Capacity Optimizer performs four main functions that, together, serve to reclaim waste, reduce cost, and improve performance in big data systems without manual intervention. Read on to learn more.

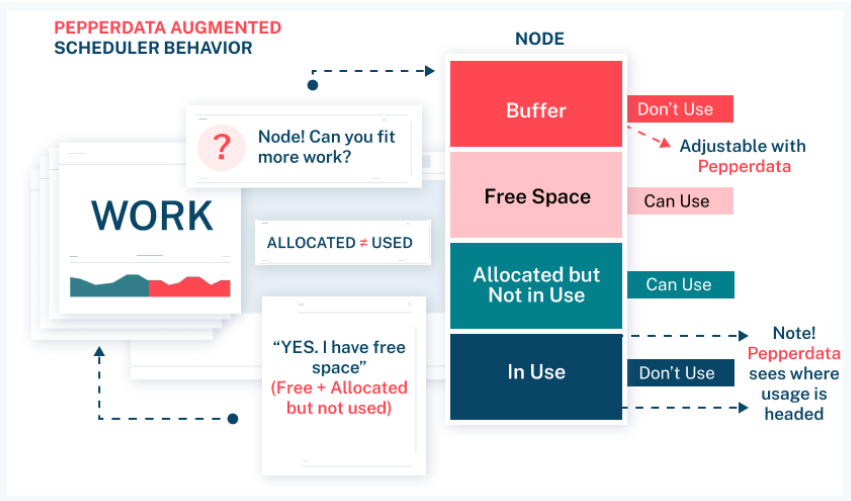
Capacity Optimizer maximizes resource utilization.

A cloud instance is maximally efficient when there is a high percentage utilization of CPU and memory running application tasks. The native YARN or Kubernetes scheduler (both in the cloud and on premises) assigns resources and runs jobs based on its imperfect knowledge of actual cluster capacity. When no capacity exists, it waits for capacity to free up or for more instances to be added before scheduling more. As a result, the cluster may experience pending workloads and a backlog of pods alongside low node utilization. That low node utilization represents resources that are being paid for and not used.

Pepperdata Capacity Optimizer enables the scheduler or cluster manager to schedule workloads based on actual resource *utilization* instead of resource *allocation*. From the scheduler's real-time view of each instance, Capacity Optimizer assesses the instantaneous load characteristics and intelligently informs the scheduler which resources are available and where nodes have resources to do more work. Capacity Optimizer makes thousands of decisions per second and implements a bin-packing type of algorithm to fill available cluster resources with waiting applications.







Suppose you have three jobs and a fixed amount of cluster resources allocated from the scheduler. Now, suppose the scheduler attempts to accommodate the first two jobs and has some resources left over, but not enough for the third job. Pepperdata Capacity Optimizer collects the near real-time metrics on exactly how many resources the first two jobs require. There is some extra buffer in those two jobs, due to factors explained above, such as overprovisioning. If that buffer is combined with the unused resources, the scheduler likely will have sufficient resources to schedule the third job.

You can imagine this as a one-dimensional game of Tetris, with all the running jobs dropping into one tight space, with contiguously aggregated free space above.

The net result: Capacity Optimizer eliminates waste and enables up to a 30 percent throughput gain with the same workload duration.

Capacity Optimizer reclaims waste.

Pepperdata Capacity Optimizer also removes capacity in the case of overprovisioned applications. This removal capability is Capacity Optimizer's advantage over simple overprovisioning. An overprovisioned application might require 10 GB of memory at peak, but that peak time may only represent 20 percent of the application's run. The other 80 percent of the time, Capacity Optimizer can reclaim those unused resources for other applications. By finding the difference between allocated and used capacity in real time, Capacity Optimizer can instruct the YARN or Kubernetes scheduler to process waiting applications, thus providing greater throughput across the cluster.

Capacity Optimizer optimizes autoscaling in the cloud.

Pepperdata Capacity Optimizer solves the problem of inefficient cloud autoscalers by enabling the scheduler or cluster manager to schedule workloads based on actual resource utilization instead of resource allocation.

Once the target resource utilization is achieved, the autoscaler adds more instances. Cloud cost optimization through Pepperdata Capacity Optimizer not only maximizes the utilization of each of the existing instances, it also ensures that the new instances are added only when the existing instances are fully utilized in an autoscaling environment. Pepperdata Capacity Optimizer manages the autoscaling behavior of the cloud platforms so that developers don't have to. In short, it achieves results that are impossible using manual methods or standard tunable parameters.

To optimize cloud costs in both Kubernetes and YARN, Capacity Optimizer does the following:

1. The autoscaler adds new instances
2. The autoscaler then waits for Capacity Optimizer to maximize resource utilization
3. Capacity Optimizer reclaims wasted resources, so the scheduler launches more applications, thus increasing the utilization of each instance
4. If the applications are still pending after the existing nodes are fully utilized, autoscaler logic kicks in and goes back to step 1

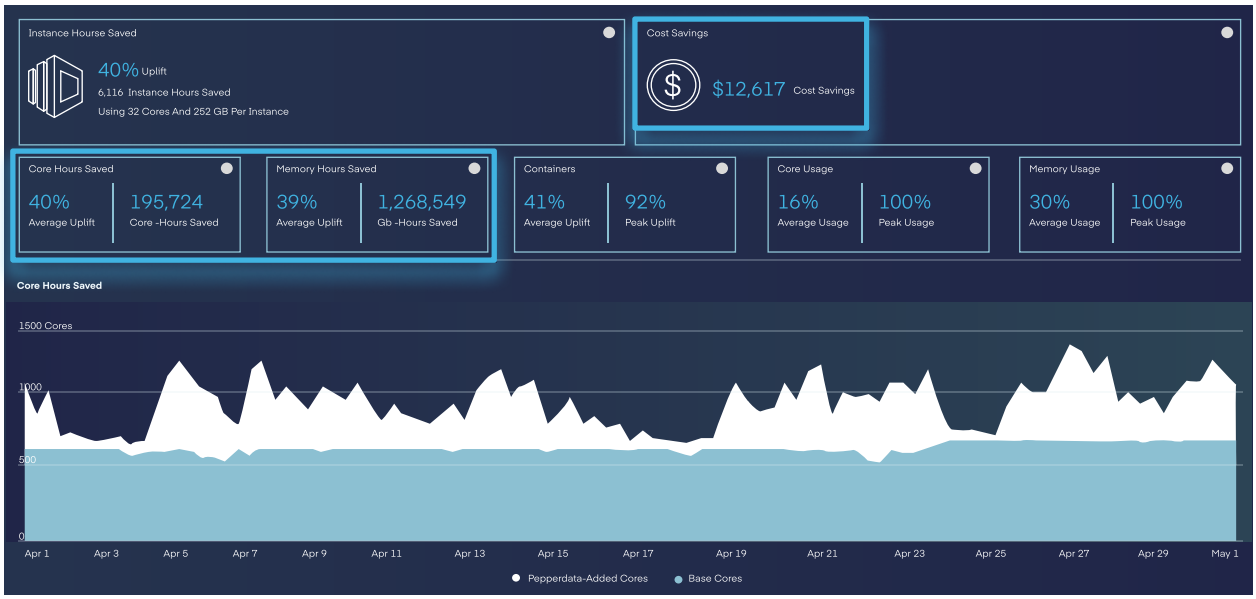
Capacity Optimizer operates autonomously and continuously.

Capacity Optimizer operates autonomously and continuously in the background as it responds to your cluster's activity in real time. Capacity Optimizer does not require manual intervention or application code changes, nor does it require you to profile your workloads ahead of time—all of which frees your developers to focus on application delivery and product innovation.

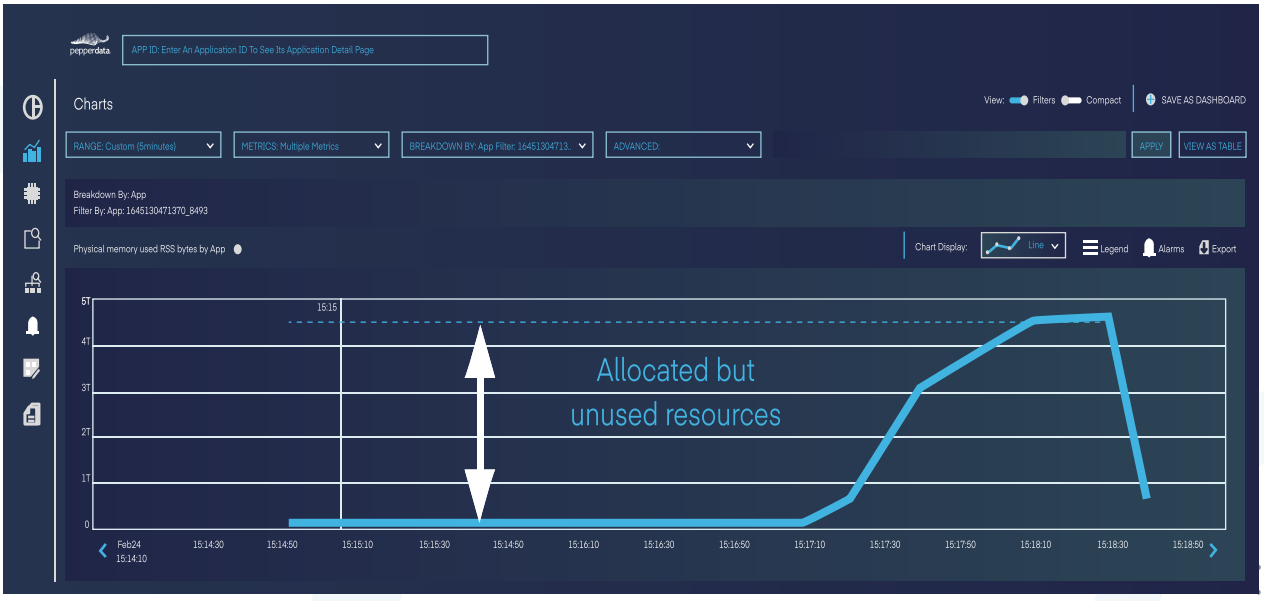
In a typical enterprise environment, Capacity Optimizer can be installed via a bootstrap script in under an hour and can deliver up to 38 percent cost reduction with the same workload duration.

Read on for examples of how Capacity Optimizer does just that in some of the world's most demanding clusters.

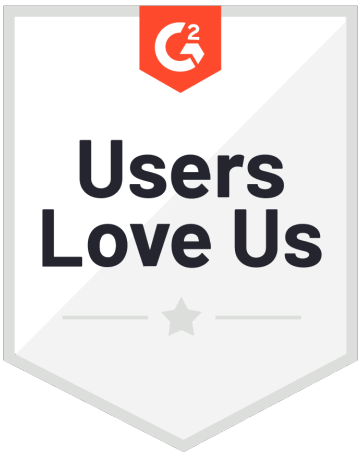
Quantify Your Savings



Visualize Your Recovered Waste



Top Enterprises Rely On Pepperdata



Enterprises that run their big data applications with Capacity Optimizer see large reductions in cost accompanied by tremendous improvements in performance. For example:

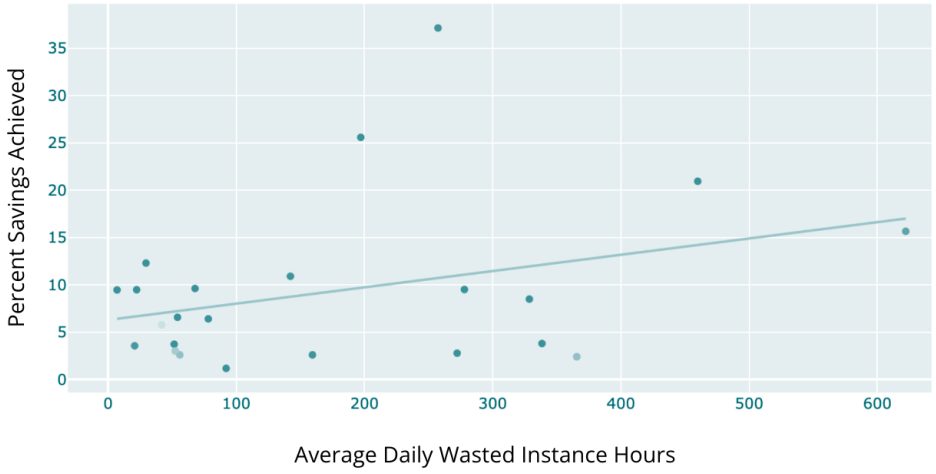
- **A Fortune 100 financial services firm** reduced server infrastructure spending by 30 percent by identifying underutilized servers and redirecting workloads to those resources. At the same time, they added the equivalent of hundreds of nodes and significantly reduced the number of servers needed to support continued growth.
- **A Fortune 100 healthcare giant** running over one thousand nodes estimates annual savings of $943K because of Capacity Optimizer.
- **A Fortune 100 customer intelligence company** running five thousand nodes estimates an annual $733K in savings from Capacity Optimizer.
- **A Fortune 5 customer** with a 10,000+ node Hadoop/YARN cluster achieves a 95 percent utilization with Capacity Optimizer.

Customers running Capacity Optimizer have more efficient resource utilization, dramatically increased throughput, and significant cost reduction, all from an autonomous service.

The above implementations also reflect a variety of big data and Kubernetes environments and technologies, including Amazon EMR, GKE, Cloudera, and Apache Spark.

Pepperdata Found up to 38 Percent Additional Savings in Clusters Where Conventional Interventions Had Been Applied

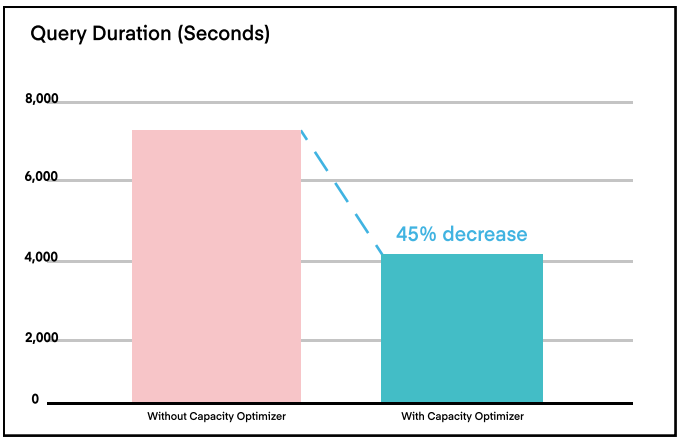
- In this chart, each dot represents a customer cluster optimized by Pepperdata
- Even after conventional interventions were applied, Pepperdata discovered instance hours were being wasted
- Pepperdata found additional savings up to 38 percent

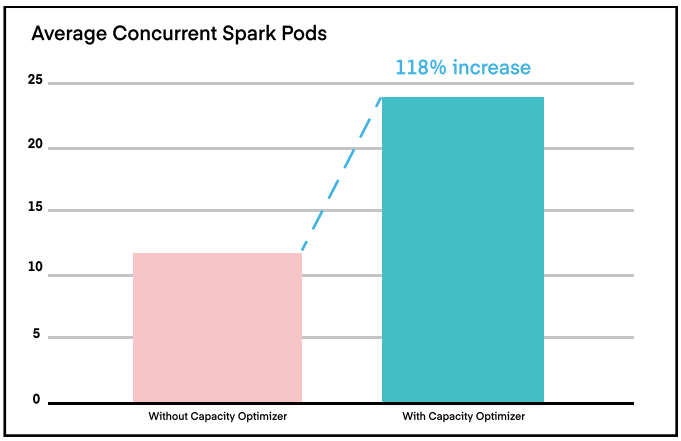


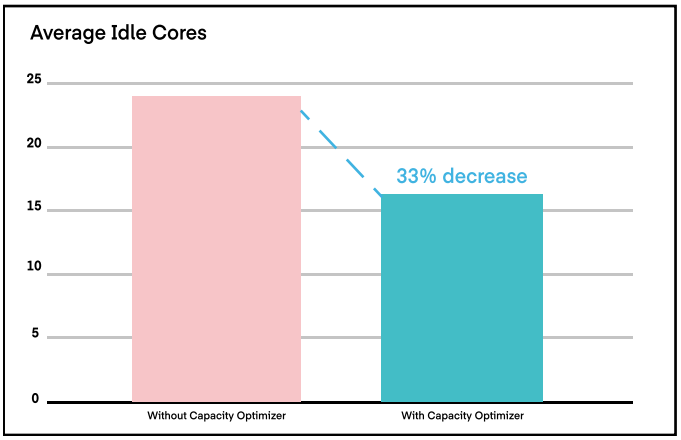
Third-Party Benchmarks Support Real-World Findings

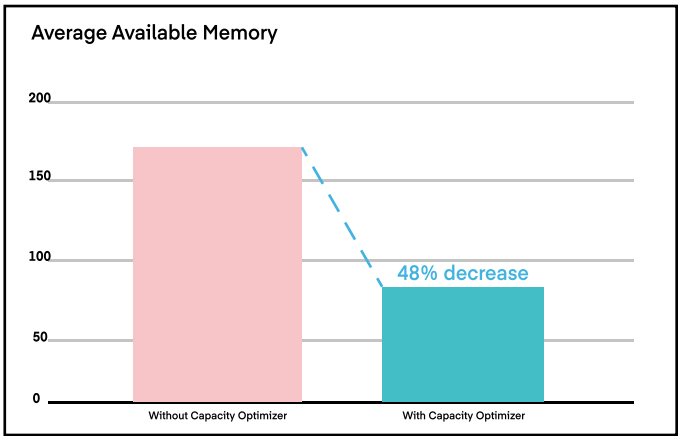
Third-party benchmark results support the real-world findings of Pepperdata's enterprise customers. On Amazon EMR, the TPC-DS benchmark runs 8 percent faster than with cloud-native custom autoscaling parameters. It shows a 38 percent decrease in instance hours, a 157 percent increase in CPU utilization, and a 38 percent increase in memory utilization, all because Capacity Optimizer is trimming the fat from the job as it runs.

(Note: TPC-DS is an industry-standard stable of sample queries that simulate an e-commerce website. Even though we used TPC-DS data and queries, this benchmark is not an official TPC-DS benchmark.)









TPC-DS on EKS with the Apache YuniKorn scheduler shows significant improvements in performance and efficiency under Capacity Optimizer: a 45 percent faster query duration, 48 percent more available memory, 38 percent more concurrently running pods, and 118 percent more concurrently running Spark pods.

DOWNLOAD THE REPORT

Capacity Optimizer Can Slash Amazon EC2 Costs By Over 50%

One Pepperdata customer is a global leader in design and manufacturing software with primary markets in the manufacturing, media, and entertainment industries. The company constantly analyzes large volumes of data using Apache Spark on Amazon EMR.

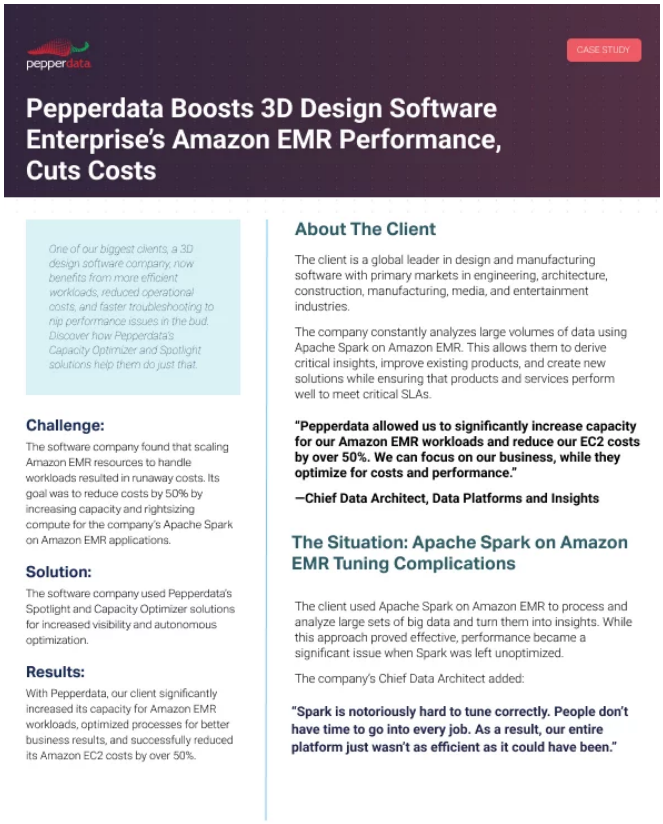
But performance was a big problem. The company's Chief Data Architect explained:

"Spark is notoriously hard to tune correctly. People don't have time to go into every job. As a result, our entire platform just wasn't as efficient as it could have been."

Amazon EMR's native autoscaling helped, but not enough. As the company's workloads grew, so did the team's Spark issues. The increased compute consumption was quickly eating through its budget. Each Amazon EMR cluster was consuming two or three times the planned capacity.

Pepperdata Capacity Optimizer solved the problem of inefficiency in the company's cloud autoscaling by enabling the scheduler or cluster manager to schedule workloads based on actual resource utilization instead of resource allocation —cutting the organization's EC2 instance cost by 50 percent.

Pepperdata Capacity Optimizer not only maximizes the utilization of each of the existing instances, it also manages the autoscaling behavior of the cloud platforms and ensures that new instances are added only when existing instances are fully utilized in an autoscaling environment.



pepperdata

CASE STUDY

Pepperdata Boosts 3D Design Software Enterprise's Amazon EMR Performance, Cuts Costs

One of our biggest clients, a 3D design software company, now benefits from more efficient workloads, reduced operational costs, and faster troubleshooting to nip performance issues in the bud. Discover how Pepperdata's Capacity Optimizer and Spotlight solutions help them do just that.

Challenge:

The software company found that scaling Amazon EMR resources to handle workloads resulted in runaway costs. Its goal was to reduce costs by 50% by increasing capacity and rightsizing compute for the company's Apache Spark on Amazon EMR applications.

Solution:

The software company used Pepperdata's Spotlight and Capacity Optimizer solutions for increased visibility and autonomous optimization.

Results:

With Pepperdata, our client significantly increased its capacity for Amazon EMR workloads, optimized processes for better business results, and successfully reduced its Amazon EC2 costs by over 50%.

About The Client

The client is a global leader in design and manufacturing software with primary markets in engineering, architecture, construction, manufacturing, media, and entertainment industries.

The company constantly analyzes large volumes of data using Apache Spark on Amazon EMR. This allows them to derive critical insights, improve existing products, and create new solutions while ensuring that products and services perform well to meet critical SLAs.

"Pepperdata allowed us to significantly increase capacity for our Amazon EMR workloads and reduce our EC2 costs by over 50%. We can focus on our business, while they optimize for costs and performance."

—Chief Data Architect, Data Platforms and Insights

The Situation: Apache Spark on Amazon EMR Tuning Complications

The client used Apache Spark on Amazon EMR to process and analyze large sets of big data and turn them into insights. While this approach proved effective, performance became a significant issue when Spark was left unoptimized.

The company's Chief Data Architect added:

"Spark is notoriously hard to tune correctly. People don't have time to go into every job. As a result, our entire platform just wasn't as efficient as it could have been."

Capacity Optimizer allowed the company to implement a more sophisticated approach to Spark performance tuning. The solution automatically optimizes the resources in its clusters and recaptures compute waste, resulting in 15 percent reduction of instance hours. With more resources available, the company can run more applications without adding additional hardware and personnel to tune them.

As the Chief Data Architect put it:

"We can focus on our business, while Pepperdata optimizes for costs and performance."

READ THE CASE STUDY

Finally, Minimize Cloud Waste and Regain Control Over Your Budget

One of the defining characteristics of Capacity Optimizer is automation. Automation has become a hallmark of the modern data stack. Developers automate the process of writing and deploying code and design the software to automate the scaling of resources as needed. Automation is necessary because cloud-native systems are not optimally efficient at providing a desired price/performance ratio, because automation begets efficiency, and because manual processes can't possibly keep up anyway.

It's the same with big data optimization. The modern data stack is too complicated for people to manage manually, so developers use automated software systems to optimize them on the fly.

Pepperdata helps some of the top companies in the world maximize the productivity of their modern big data initiatives. Enterprises hire developers to bridge the gap between software applications and hardware, not to tune code to the particulars of a complex execution environment. Unlike passive observability tools, which rely on recommendations and manual tuning, Pepperdata is an autonomous, scalable solution that helps IT to balance big data performance goals and resource availability, eliminate manual tuning, and manage spend for improved ROI.

Ready to Get Started?

Could your organization benefit from pressing an "easy button" to reclaim waste and reduce cloud costs or extend the life of your on-prem investment? We'd welcome the opportunity to provide it to you.

Capacity Optimizer requires no manual changes to your application code or platform settings and can be installed via a bootstrap script in under an hour in most enterprise environments.

Please get in touch with us for a free trial of Pepperdata, or email us at any time with your questions at sales@pepperdata.com.

READY TO GET STARTED?

Could your organization benefit from pressing an "easy button" to reduce cloud costs or extend the life of your on-prem investment? We'd welcome the opportunity to provide it to you. Please get in touch with us for a free trial of Pepperdata or email us at any time with your questions at sales@pepperdata.com.

pepperdata

3945 Freedom Circle, Suite 920 Santa Clara, CA 9505

Start a Free Trial
www.pepperdata.com/trial

Send an Email
eval@pepperdata.com