pepperdata

BENCHMARK

Benchmark Shows Pepperdata Decreases Instance Hours/Cost by 38% on Amazon EMR

Background: Benchmarking Big Data on AWS

This report covers Pepperdata's initial benchmarking work in 2021 based on TPC-DS, an industry-standard big data analytics Decision Support framework from the Transaction Processing Performance Council. While not an official audited benchmark as defined by TPC, this work demonstrates the performance, efficiency, and cost improvements Pepperdata Capacity Optimizer delivers on Amazon EMR in addition to the benefits of the standard AWS Custom Auto Scaling Policy.

Pepperdata ran the benchmark "out of the box"and did not modify or recompile data using any special libraries.

There were to two groups of iterations for the benchmark, with results averaged across the iterations for a group:

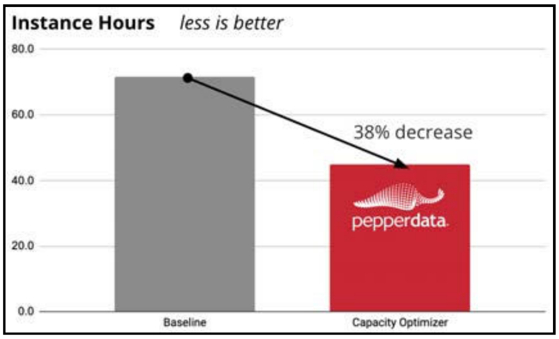
- **Baseline:** Amazon EMR with Custom Auto Scaling rules (min: 5, max: 30 nodes)
- **Optimized:** Same settings as Baseline but with Pepperdata Capacity Optimizer enabled

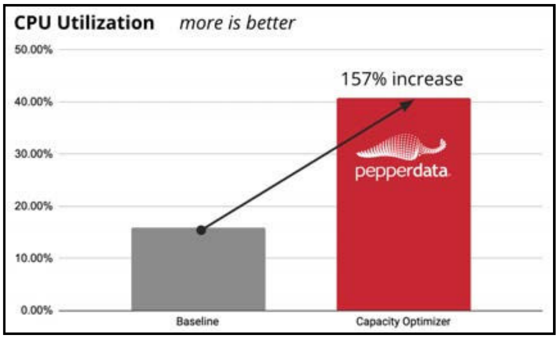
Each iteration consisted of three distinct TPC-DS workload disciplines: Database Load, Query Run, and Data Maintenance. The query run was executed twice, once before and once after the data maintenance step. Each query run executed the same 103 query templates with different variables in permuted order, thereby simulating a workload of multiple concurrent users accessing the system.

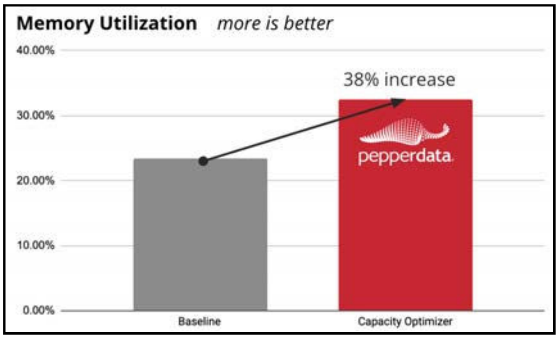
Key Findings

This report highlights three groups of findings that demonstrate that Pepperdata Capacity Optimizer can automatically:

- **Decrease instance hours and thus reduce cost** by 38 percent
- **Optimize resource utilization**, as measured by a 157 percent increase in CPU utilization and 38 percent increase in memory utilization







Instance Hour Results

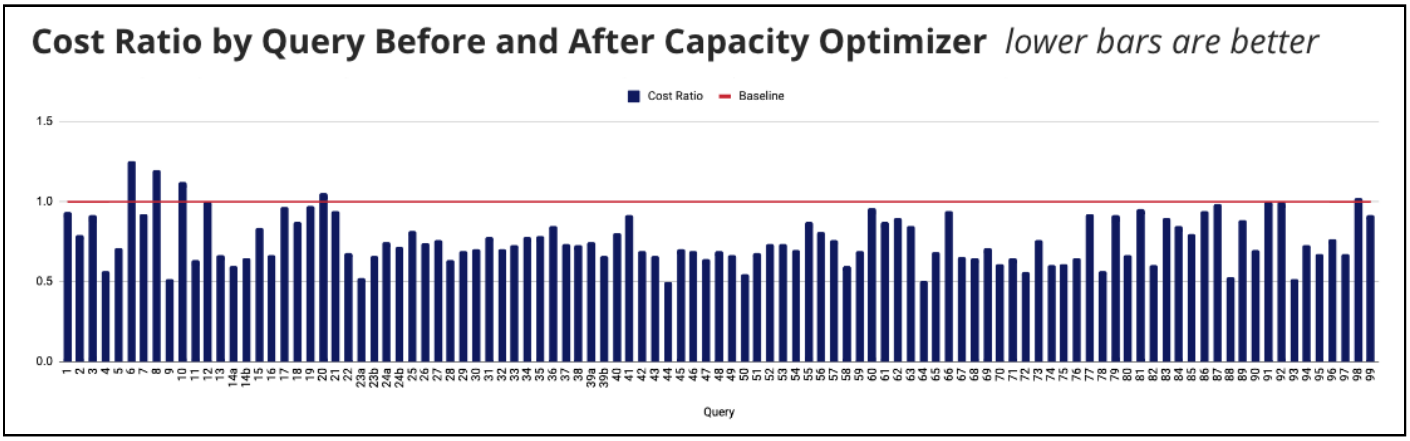
The instance hour results warrant special consideration because a **decrease in instance hours correlates directly to reduced cost.** The 38 percent decrease in instance hours that Pepperdata achieved meant that the cost to run this set of queries **cost 38 percent less when Capacity Optimizer was enabled**.

Resource Utilization Results

On average, Capacity Optimizer increased both CPU utilization by **157 percent** and memory utilization by **38 percent** when compared baseline iterations without Pepperdata optimizations for the TPC-DS workload. Utilizing CPU and memory resources more efficiently can enhance overall performance by enabling applications to run more smoothly and respond faster, as well as enable more applications to run concurrently with fewer performance bottlenecks.

Cost Ratio Results

Pepperdata compared the cost of the two options for each of the 103 queries. In the following chart, the baseline cost line represents the average cost of the queries with TPC-DS data using the AWS Custom Autoscaling. Any queries where Capacity Optimizer-provided savings are represented by bars which run below the baseline. **More than 90 percent of the queries which occurred while running Capacity Optimizer resulted in additional savings versus running the AWS Custom Auto Scaling policy alone.**

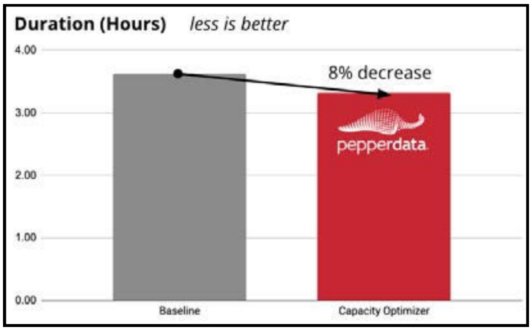


Cost Ratio by Query Before and After Capacity Optimizer *lower bars are better*

The five queries where Capacity Optimizer did not add additional savings (those above the red line)—TPC-DS Query 6, TPC-DS Query 8, TPC-DS Query 10, TPC-DS Query 20, and TPC-DS Query 98— were the simpler and less complicated queries. Pepperdata excelled in the most complicated queries that reflect the most demanding real-world environments, comprising fully 90 percent of the workloads.

Duration Results

Although overall duration was not a primary metric in evaluating the effectiveness of Capacity Optimizer, we did observe an approximate 8 percent decrease in the overall runtime of the entire query suite, with the overall duration decreasing from 3.62 hours to 3.33 hours.



Conclusion

Using the industry-standard benchmark dataset TPC-DS, Pepperdata demonstrated that Capacity Optimizer provides an uplift to the Amazon AWS Custom Auto Scaling Policy in both CPU and memory utilization while decreasing instance hours and overall time to run the entire suite of 103 queries.

As more companies migrate big data workloads to the cloud, these findings have important implications for cost and resource management. A recent survey conducted by Pepperdata identified “significant or unexpected spend” as a primary challenge of organizations moving to the cloud and adopting Kubernetes. By providing real-time, autonomous cost optimization, Pepperdata Capacity Optimizer enables cloud workloads to run more efficiently, resulting in substantial cost and resource savings. This makes the cloud an even more attractive and viable option for large-scale workloads.

For information on the methodology and configurations used and referenced in this report, please contact us at info@pepperdata.com.

100% ROI
GUARANTEE
pepperdata

Pepperdata installs in under 30 minutes in most enterprise environments. We guarantee a minimum of 100% ROI, with a typical ROI between 100% and 660%.

About Pepperdata

Pepperdata is the only cost optimization solution that delivers up to 47% greater cost savings—continuously and in real-time—on Amazon EMR and EKS with no application changes or manual tuning. Our customers include the largest, most complex, and highly-scaled clusters in the world, at top enterprises such as Citibank, Autodesk, Royal Bank of Canada, and those in the Fortune 5. For more information, visit pepperdata.com.

Pepperdata, Inc.
530 Lakeside Drive
Suite 170
Sunnyvale, CA 94085

Start a Free PoV
www.pepperdata.com

Send an Email
eval@pepperdata.com