CASE STUDY

How OptScale allowed the company with an $80M cloud bill to run ML experiments with optimal performance and reduce an infrastructure cost by 37%

EXECUTIVE SUMMARY

Due to a significant number of ML processes launched by hundreds of ML engineers, the mobile advertising broker company (The Company, with more than 800 employees, providing the leading mobile advertising platform, has a complex IT infrastructure and high cloud costs. Leveraging the AWS platform for hundreds of ML models, the company spent over $80M annually on a cloud environment.

OptScale helped reduce AWS cloud costs by 37% in four months by optimizing ML/AI workload performance, organizing experiment tracking, improving ML teams' KPI and delivering the company's cloud usage and cost transparency.

THE GOAL

The Company aimed to empower the MLOps process by implementing MLOps and FinOps methodology, providing complete transparency of ML model training process with a leaderboard and experiment tracking and optimizing ML experiment performance and cost.

MAIN CHALLENGES

Running hundreds of ML experiments on a daily basis, ML teams faced the following challenges:

- Lack of automated and efficient instruments for ML/AI model training tracking and profiling/instrumentation.

ML/AI model training is a complex process that depends on a defined hyperparameter set, hardware, or cloud resource usage. Monitoring and comparing key metrics and indicators against established benchmarks or thresholds enable gaining profound insights and enhancing the ML/AI profiling process.

- Limited transparency throughout the ML lifecycle.

Without sufficient transparency into the ML process, it became challenging for the company to determine bottlenecks in ML model training and select the optimal configuration of cloud resources. The lack of visibility hinders the ability to maximize ML/AI training resource utilization and outcome of experiments and accurately plan and forecast resource requirements, leading to overprovisioning or underprovisioning cloud resources.

- Identifying optimization scenarios for improving performance and cloud bill reduction.

ML models often require complex and significant cloud infrastructure for training and inference. Inefficient ML model and experiment management mechanisms resulted in increased resource costs and longer processing times due to bottlenecks in specific resources like GPU, IO, CPU, or RAM. Without proper monitoring, the company faced challenges in identifying bottlenecks, performance issues, or areas for improvement.

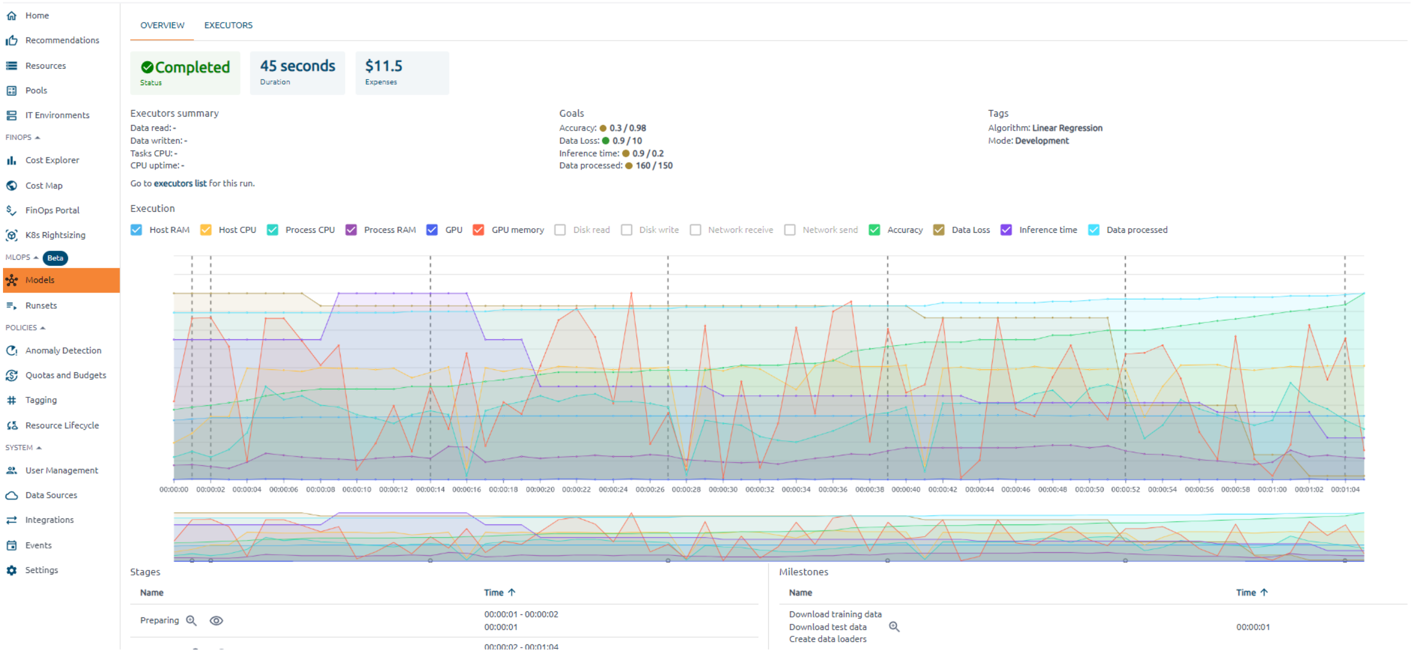
THE SOLUTION

Hystax OptScale allowed the mobile advertising broker to improve their ML process by:

1. Providing **ML model leaderboards** gives an ML team full transparency across the ML model metrics and performance, which helps find the optimal combinations of parameters and the best result of ML experiments.
2. **ML task profiling with an in-depth analysis** of performance metrics and experiment tracking. With OptScale, ML/AI engineering teams get an instrument for tracking and profiling ML/AI model training and other relevant tasks. OptScale collects a holistic set of performance and model-specific metrics, which provide performance enhancement and cost optimization recommendations for ML/AI experiments or production tasks.

3 Delivering **optimization recommendations** enabled the company to save up to 37% of the monthly AWS cloud bill and gain infrastructure usage transparency. The recommendation included using optimal Reserved Instances and Savings Plans, rightsizing, detecting unused resources, cost allocation, and others.

4 **Runsets** - automated run of a number of experiments with configurable datasets, hyperparameter ranges, and model versions. Runsets enabled running experiments parallel with various input parameters and identifying the most efficient ML/AI model training results.



Completed
Status

45 seconds
Duration

$11.5
Expenses

Executors summary
Data read: -
Data written: -
Tasks CPU: -
CPU uptime: -

Go to **executors list** for this run.

Goals
Accuracy: ● **0.3 / 0.98**
Data Loss: ● **0.9 / 10**
Inference time: ● **0.9 / 0.2**
Data processed: ● **160 / 150**

Tags
Algorithm: **Linear Regression**
Mode: **Development**

Execution

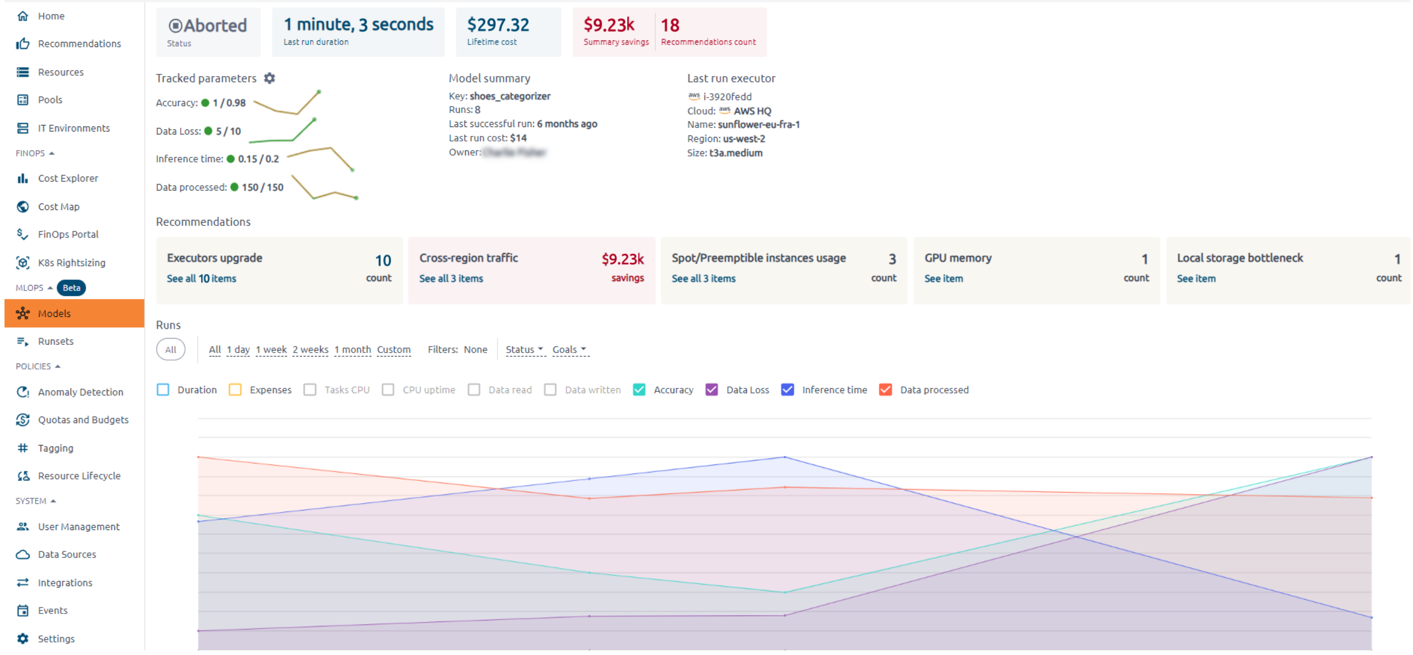
☑ Host RAM ☑ Host CPU ☑ Process CPU ☑ Process RAM ☑ GPU ☑ GPU memory □ Disk read □ Disk write □ Network receive □ Network send ☑ Accuracy ☑ Data Loss ☑ Inference time ☑ Data processed

Stages

Name	Time ↑
Preparing	00:00:01 - 00:00:02 00:00:01
	00:00:02 - 00:01:04

Milestones

Name	Time ↑
Download training data Download test data Create data loaders	00:00:01



Aborted
Status

1 minute, 3 seconds
Last run duration

$297.32
Lifetime cost

$9.23k
Summary savings

18
Recommendations count

Tracked parameters
Accuracy: ● **1 / 0.98**
Data Loss: ● **5 / 10**
Inference time: ● **0.15 / 0.2**
Data processed: ● **150 / 150**

Model summary
Key: **shoes_categorizer**
Runs: **8**
Last successful run: **6 months ago**
Last run cost: **$14**
Owner: [redacted]

Last run executor
aws i-3920fedd
Cloud: aws **AWS HQ**
Name: **sunflower-eu-fra-1**
Region: **us-west-2**
Size: **t3a.medium**

Recommendations

Executors upgrade	10	See all **10** items	count
Cross-region traffic	$9.23k	See all 3 items	savings
Spot/Preemptible instances usage	3	See all 3 items	count
GPU memory	1	See item	count
Local storage bottleneck	1	See item	count

Runs

All | All 1 day 1 week 2 weeks 1 month Custom | Filters: None | Status ▾ Goals ▾

□ Duration □ Expenses □ Tasks CPU □ CPU uptime □ Data read □ Data written ☑ Accuracy ☑ Data Loss ☑ Inference time ☑ Data processed

RESULTS

OptScale allowed the company to run ML experiments with optimal performance, reduced infrastructure costs, and **improve their KPIs (key innovation index)**.

Using Hystax OptScale ML team multiplied the number of ML/AI experiments running in parallel, maximized ML/AI training resource utilization and outcome of experiments, reduced model training time, and minimized cloud costs. The solution enabled ML/AI engineers to run automated experiments based on datasets and hyperparameter conditions within a defined infrastructure budget.

OptScale enabled ML teams to manage the lifecycle of models and experiment results through simplified cloud management and enhanced user experience.

ABOUT HYSTAX

Hystax develops OptScale, an MLOps & FinOps open source platform that optimizes performance and IT infrastructure cost by analyzing cloud usage, profiling and instrumentation of applications, ML/AI tasks, and cloud PaaS services, and delivering tangible optimization recommendations. The tool aims to find performance bottlenecks, optimize cloud spend and give a complete picture of utilized cloud resources and their usage details.
The platform can be used as a SaaS or deployed from source code; it is optimized for ML/AI teams but works with any workload.

Github project: https://github.com/hystax/optscale

OptScale live demo: https://my.optscale.com/live-demo