Hystax OptScale

MLOps open source platform

Multiply a number of ML/AI experiments
with minimal cloud costs

hystax.com | info@hystax.com | 1250 Borregas Avenue Sunnyvale, CA 94089

Hystax

Founded in 2016,
customers in 48 countries

Customers: Airbus, Nutanix,
Orange, Nokia, DHL, Burger King

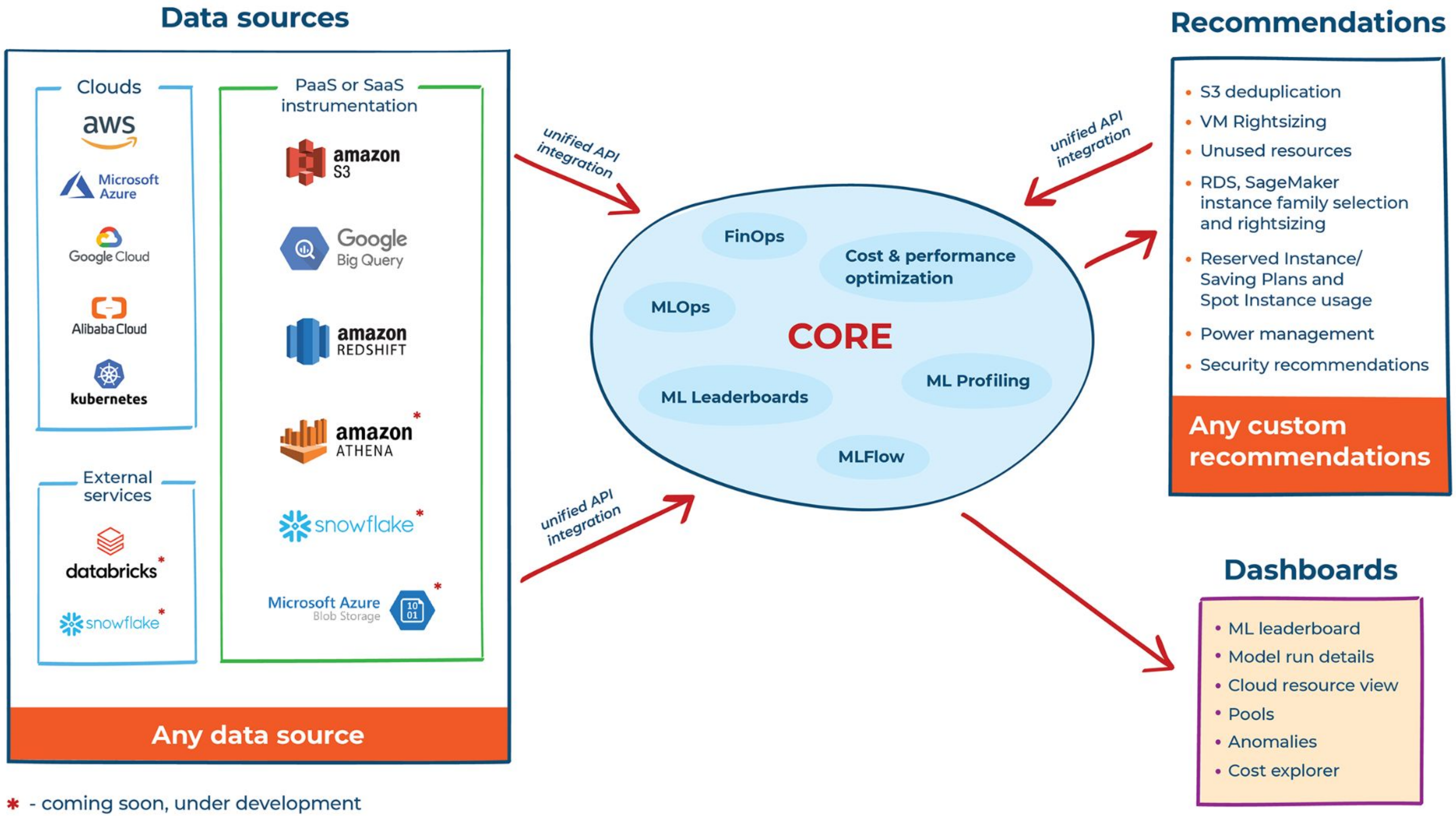
OptScale use cases

MLOps

FinOps & cloud cost optimization

Open source under Apache 2.0: https://github.com/hystax/optscale

OptScale schema



Data sources

Clouds

External services

PaaS or SaaS instrumentation

Any data source

CORE

FinOps

Cost & performance optimization

MLOps

ML Leaderboards

ML Profiling

MLFlow

Recommendations

- S3 deduplication
- VM Rightsizing
- Unused resources
- RDS, SageMaker instance family selection and rightsizing
- Reserved Instance/ Saving Plans and Spot Instance usage
- Power management
- Security recommendations

Any custom recommendations

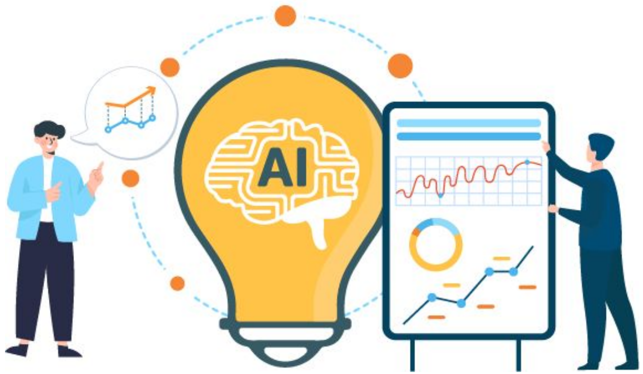
Dashboards

- ML leaderboard
- Model run details
- Cloud resource view
- Pools
- Anomalies
- Cost explorer

* - coming soon, under development

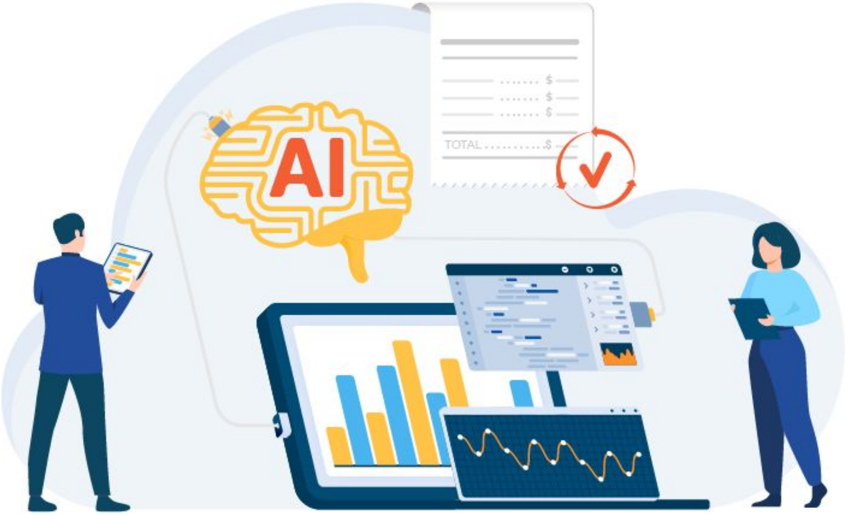
MLOps

- Team and individual ML engineer progress observability
- ML/AI task profiling, bottleneck identification
- PaaS or any external service instrumentation
- Optimization recommendations
- Runsets to automatically scale a number of experiments

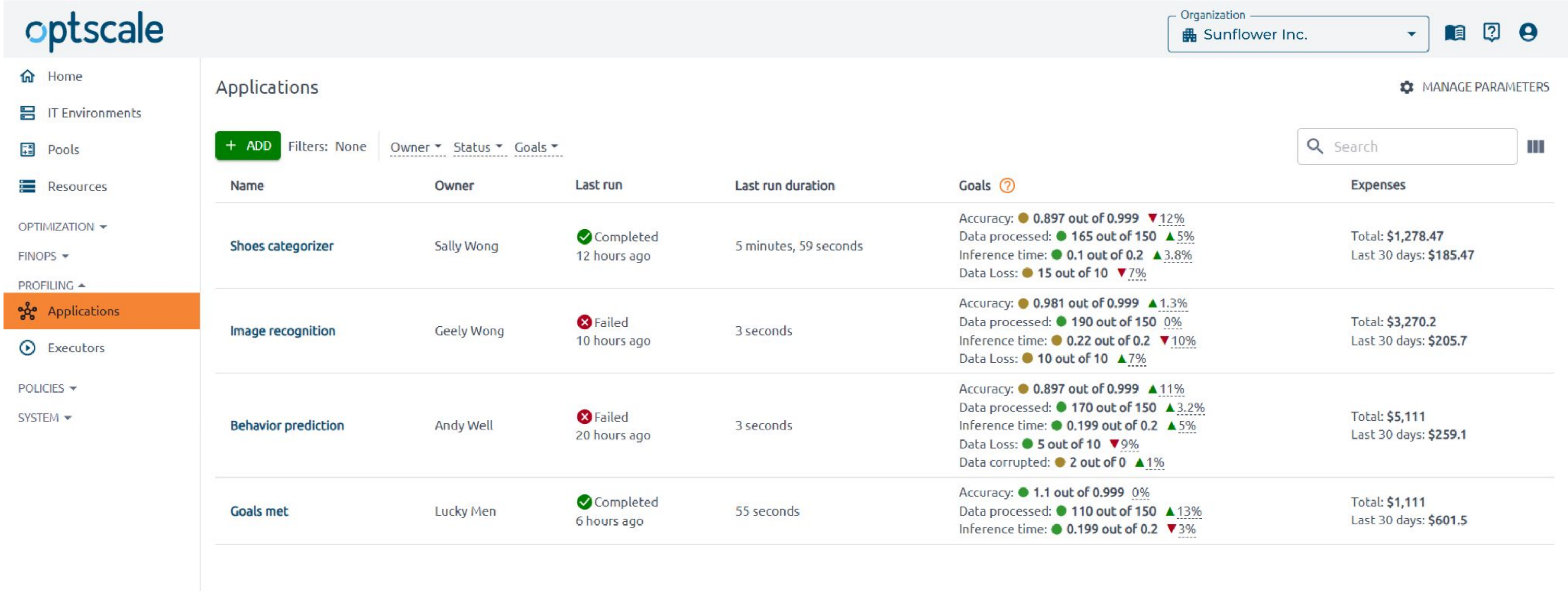


ML R&D status observability

- List of models with goals status and active recommendations
- Tracking the number and quality of experiments run by a team
- Cost of an overall model and individual experiments



ML R&D status observability



Applications

Name	Owner	Last run	Last run duration	Goals	Expenses
Shoes categorizer	Sally Wong	✓ Completed 12 hours ago	5 minutes, 59 seconds	Accuracy: ● 0.897 out of 0.999 ▼12% Data processed: ● 165 out of 150 ▲5% Inference time: ● 0.1 out of 0.2 ▲3.8% Data Loss: ● 15 out of 10 ▼7%	Total: $1,278.47 Last 30 days: $185.47
Image recognition	Geely Wong	✗ Failed 10 hours ago	3 seconds	Accuracy: ● 0.981 out of 0.999 ▲1.3% Data processed: ● 190 out of 150 0% Inference time: ● 0.22 out of 0.2 ▼10% Data Loss: ● 10 out of 10 ▲7%	Total: $3,270.2 Last 30 days: $205.7
Behavior prediction	Andy Well	✗ Failed 20 hours ago	3 seconds	Accuracy: ● 0.897 out of 0.999 ▲11% Data processed: ● 170 out of 150 ▲3.2% Inference time: ● 0.199 out of 0.2 ▲5% Data Loss: ● 5 out of 10 ▼9% Data corrupted: ● 2 out of 0 ▲1%	Total: $5,111 Last 30 days: $259.1
Goals met	Lucky Men	✓ Completed 6 hours ago	55 seconds	Accuracy: ● 1.1 out of 0.999 0% Data processed: ● 110 out of 150 ▲13% Inference time: ● 0.199 out of 0.2 ▼3%	Total: $1,111 Last 30 days: $601.5

ML/AI profiling & optimization

- ML/AI model training tracking and profiling, inside and outside metrics collection
- CPU/RAM/GPU/Disk IO correlation tracking
- Minimal cloud cost for ML/AI experiments and development by utilizing Reserved Instances/Savings Plans and dozens of optimization scenarios

Supported platforms:

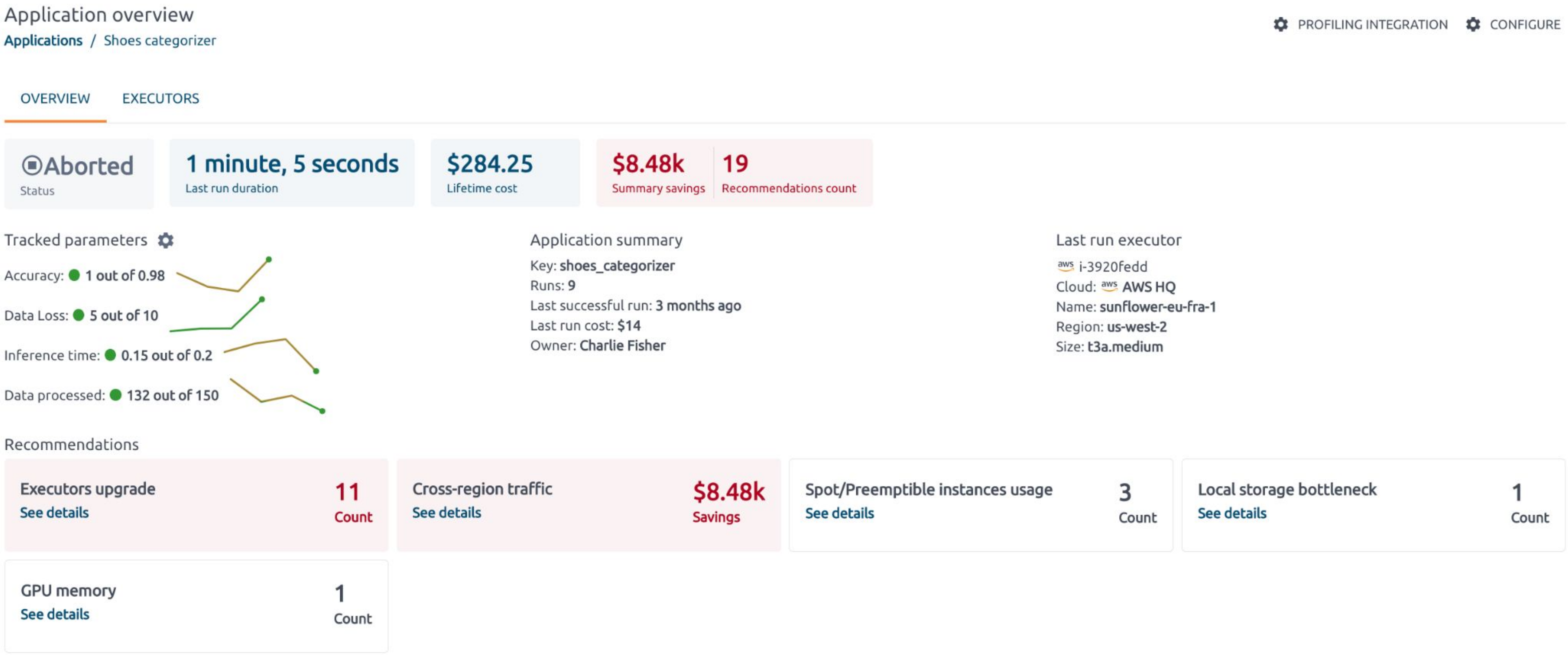
aws Microsoft Azure Google Cloud Alibaba Cloud kubernetes

databricks PyTorch Kubeflow APACHE Spark TensorFlow

ML/AI optimization recommendations

- Utilizing Reserved/Spot Instances and Savings Plans
- Rightsizing and instance family migration
- Detecting CPU, GPU, RAM, and IO bottlenecks
- Cross-regional traffic
- Experiment/run comparison

ML/AI profiling & optimization



Application overview

Applications / Shoes categorizer

PROFILING INTEGRATION   CONFIGURE

OVERVIEW   EXECUTORS

Aborted
Status

1 minute, 5 seconds
Last run duration

$284.25
Lifetime cost

$8.48k
Summary savings

19
Recommendations count

Tracked parameters

Accuracy: ● 1 out of 0.98

Data Loss: ● 5 out of 10

Inference time: ● 0.15 out of 0.2

Data processed: ● 132 out of 150

Application summary

Key: shoes_categorizer
Runs: 9
Last successful run: 3 months ago
Last run cost: $14
Owner: Charlie Fisher

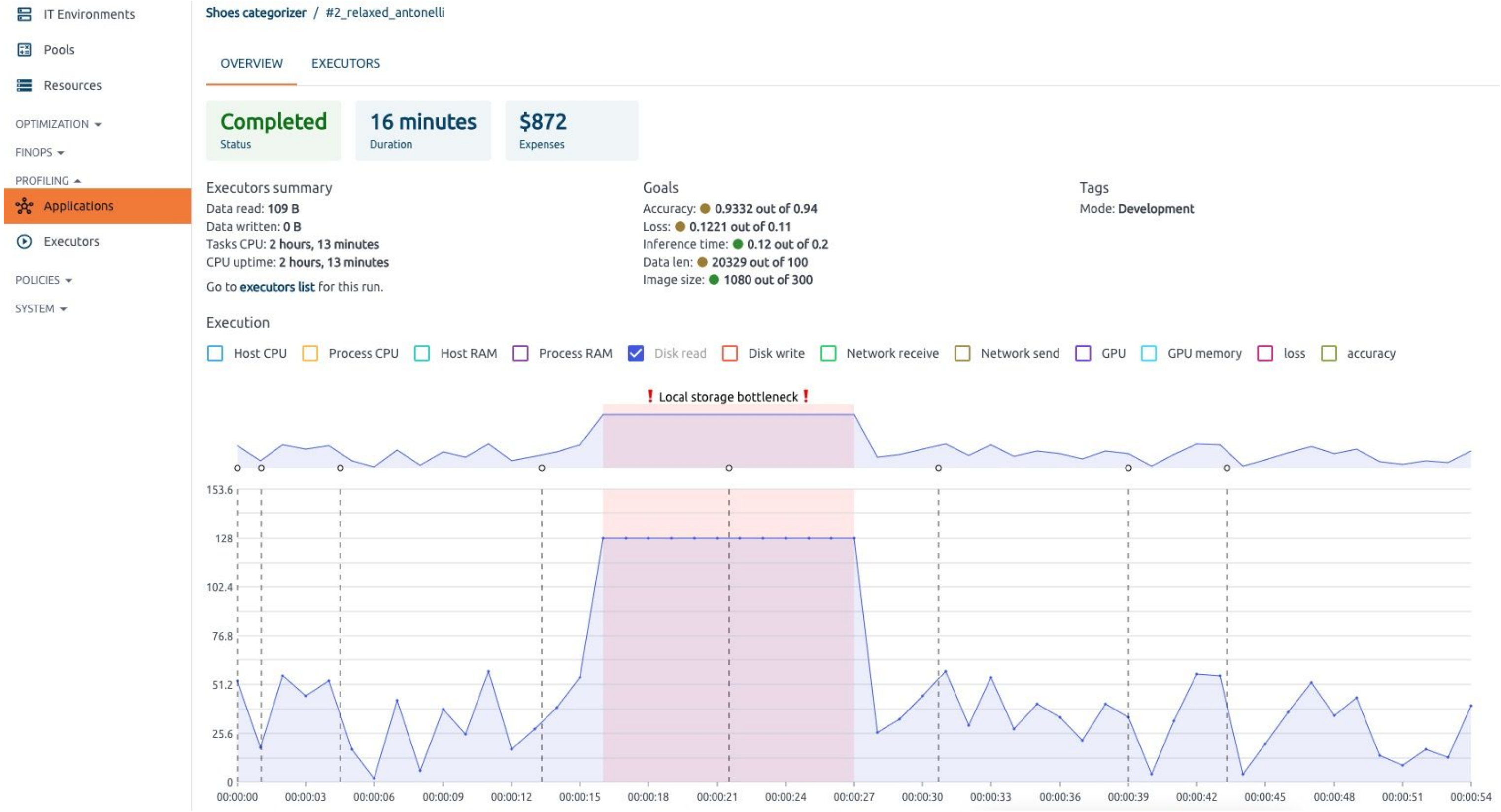
Last run executor

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

Recommendations

Executors upgrade See details	11 Count
Cross-region traffic See details	$8.48k Savings
Spot/Preemptible instances usage See details	3 Count
Local storage bottleneck See details	1 Count
GPU memory See details	1 Count

ML/AI profiling & optimization



Shoes categorizer / #2_relaxed_antonelli

Completed
Status

16 minutes
Duration

$872
Expenses

Executors summary

Data read: **109 B**
Data written: **0 B**
Tasks CPU: **2 hours, 13 minutes**
CPU uptime: **2 hours, 13 minutes**

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

Goals

Accuracy: ● **0.9332 out of 0.94**
Loss: ● **0.1221 out of 0.11**
Inference time: ● **0.12 out of 0.2**
Data len: ● **20329 out of 100**
Image size: ● **1080 out of 300**

Tags

Mode: **Development**

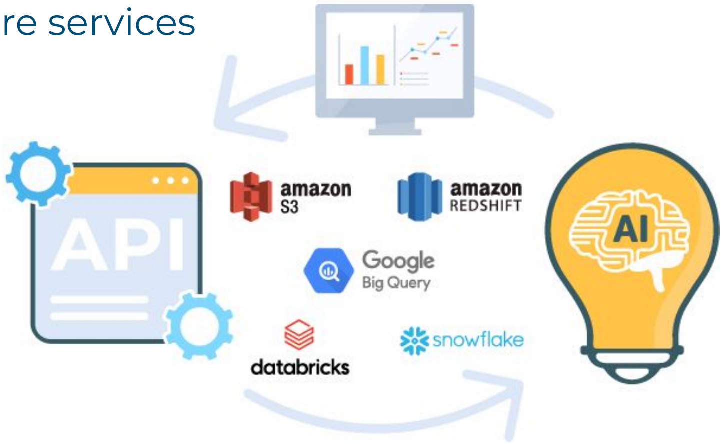
Execution

□ Host CPU □ Process CPU □ Host RAM □ Process RAM ☑ Disk read □ Disk write □ Network receive □ Network send □ GPU □ GPU memory □ loss □ accuracy

❗ Local storage bottleneck ❗

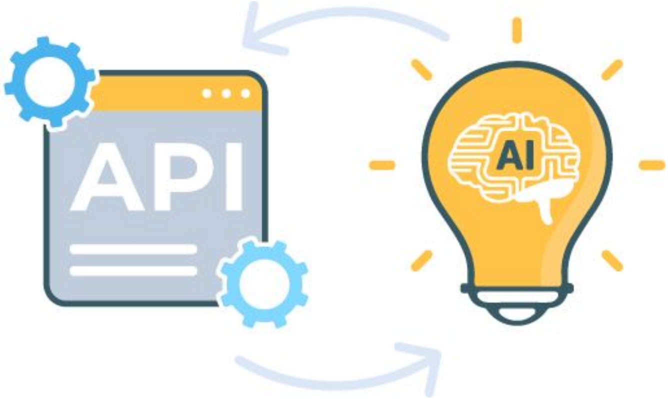
PaaS or any external service instrumentation

- Cost, performance, and output details of any API call to PaaS or an external service
- Metrics tracking and visualization
- Performance and cost optimization of API calls
- Cross-regional traffic
- S3, Redshift, BigQuery - ready, unified way to add more services

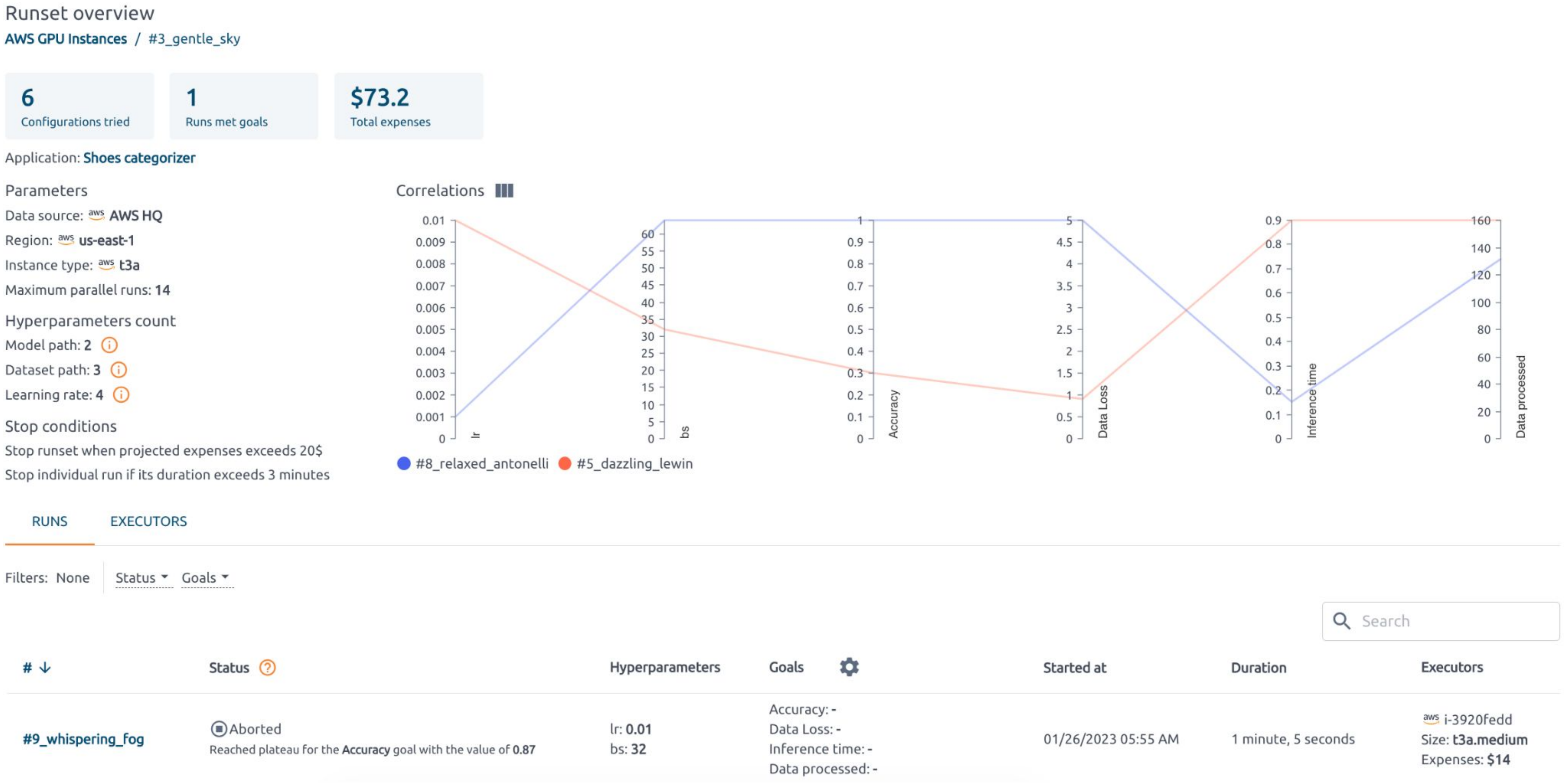


Runsets

- Automated run of a number of experiments with configurable datasets, hyperparameter ranges and model versions
- Optimal hardware with cost-efficient usage of Spot, Reserved Instances / Savings Plans
- Configurable experiment goals and success criteria
- Various complete/abort conditions - take first successful, complete all
- Integrated profiling to identify bottlenecks



Runsets



Runset overview

AWS GPU Instances / #3_gentle_sky

6 Configurations tried

1 Runs met goals

$73.2 Total expenses

Application: **Shoes categorizer**

Parameters

Data source: aws **AWS HQ**

Region: aws **us-east-1**

Instance type: aws **t3a**

Maximum parallel runs: **14**

Hyperparameters count

Model path: **2**

Dataset path: **3**

Learning rate: **4**

Stop conditions

Stop runset when projected expenses exceeds 20$

Stop individual run if its duration exceeds 3 minutes

Correlations

● #8_relaxed_antonelli ● #5_dazzling_lewin

RUNS EXECUTORS

Filters: None Status Goals

Search

# ↓	Status	Hyperparameters	Goals	Started at	Duration	Executors
#9_whispering_fog	Aborted Reached plateau for the **Accuracy** goal with the value of **0.87**	lr: **0.01** bs: **32**	Accuracy: - Data Loss: - Inference time: - Data processed: -	01/26/2023 05:55 AM	1 minute, 5 seconds	aws i-3920fedd Size: **t3a.medium** Expenses: **$14**

Roadmap

- Cost plugin for MLflow, WanDB, and neptune.ai
- Integration with Optuna to optimize Reserved Instance and other hardware parameter usage
- Model versioning
- Better hardware selection recommendations based on usage patterns and algorithms

FINOPS & CLOUD COST OPTIMIZATION

FinOps and cost management

- Forecast and monitor an IT infrastructure cost
- Identify wastage and optimize IT expenses
- Bring resource / application / service observability
- IT asset management
- Set TTL and budget constraints
- Establish a long-term FinOps process by engaging engineering teams

Supported platforms: aws Microsoft Azure Google Cloud Alibaba Cloud kubernetes

OptScale vs cloud-native cost explorer

- Cloud resource visibility and filtering across all the clouds, accounts and regions
- Dozens of optimization scenarios not supported by clouds incl. one of the best rightsizing engines
- Cost allocation not just by tags but other properties
- Geo and network traffic map
- TTL rules and budget constraints
- FinOps: OptScale is built for engineers to be responsible for their cloud resources

Contacts

☎ +1 628 251 1280

🌐 hystax.com

✉ info@hystax.com

📍 1250 Borregas Avenue, Sunnyvale, CA, 94089