Hystax OptScale

FinOps and MLOps open source platform

Run ML/AI or any type of workload with optimal performance and infrastructure cost

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

Hystax

Founded in 2016,
customers in 48 countries

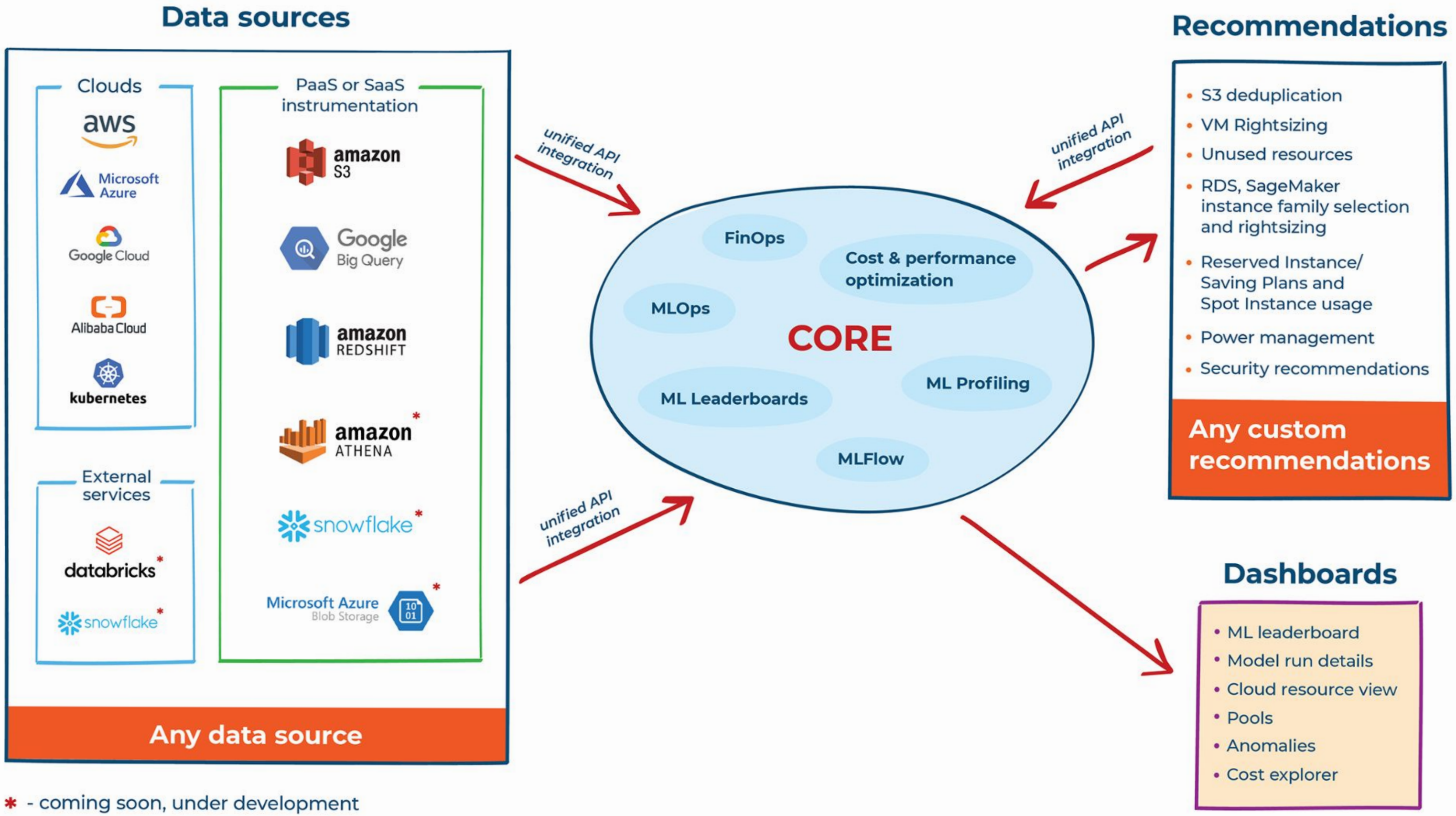
Customers: Airbus, Nutanix, Orange, Nokia, DHL, Burger King

OptScale use cases

FinOps & cloud cost optimization

MLOps

OptScale schema



Data sources

Clouds

- aws
- Microsoft Azure
- Google Cloud
- Alibaba Cloud
- kubernetes

External services

- databricks*
- snowflake*

PaaS or SaaS instrumentation

- amazon S3
- Google Big Query
- amazon REDSHIFT
- amazon ATHENA*
- snowflake*
- Microsoft Azure Blob Storage*

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

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

Cloud cost management vs FinOps

Cloud cost management:

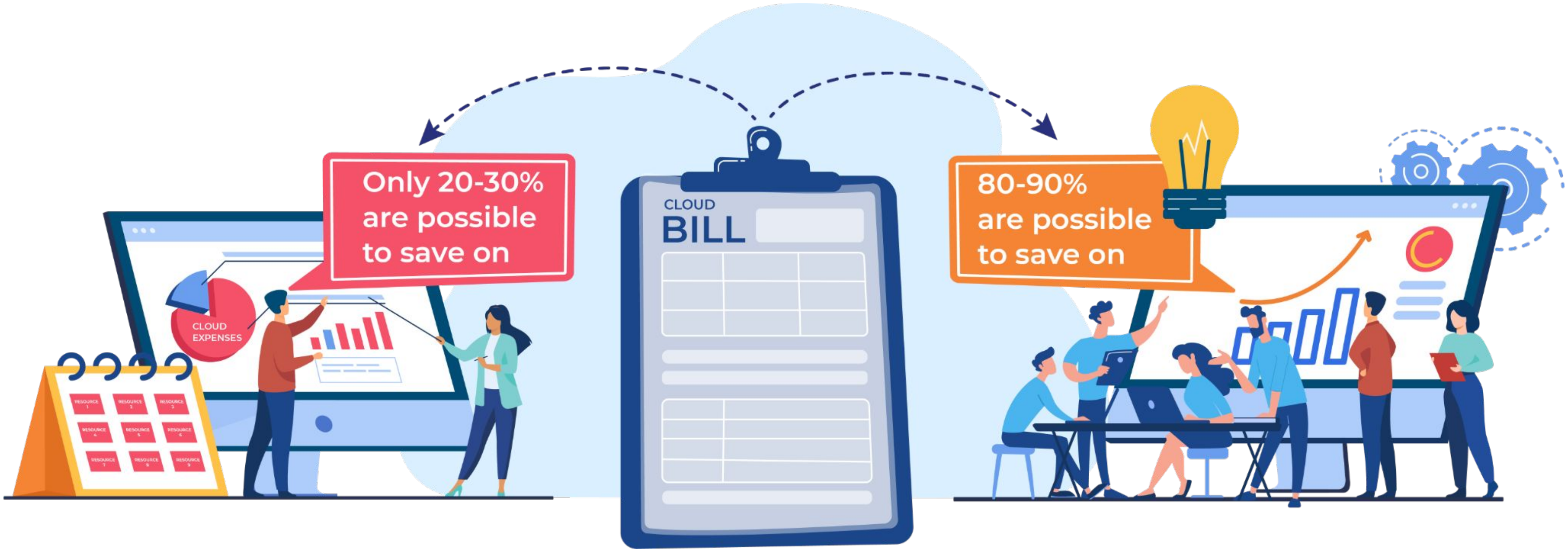
- Focused on an IT guy who needs to chase R&D teams to tag and rightsize resources, remove unused
- Gives a report to help in a short-term, in a few months issues return
- R&D team is disconnected from the cost-saving process and has no responsibility

FinOps:

- Focused on the whole FinOps team including engineers who generate the majority of costs
- Builds a cost-saving long-term process by engaging and educating the team
- IT guys are responsible for building best practices; engineers - for their own resources and TTLs; OptScale - for educating teams and delivering best practices

Cloud cost management

FinOps



Only 20-30% are possible to save on

80-90% are possible to save on

Cloud cost management solutions are built only for a few IT guys responsible for cost savings but they have limited power and influence on R&D teams

FinOps involves company's executives, financial and engineering teams in cost-saving processes

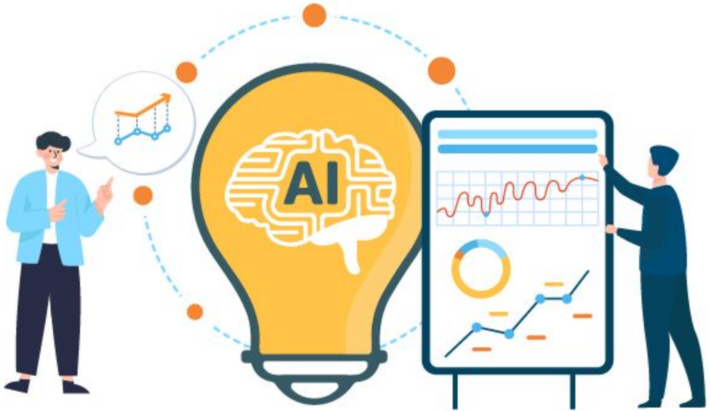
Hystax runs ‘FinOps and MLOps in Practice’, a leading FinOps and MLOps community with **9K+** members

https://finopsinpractice.org

MLOPS: ML/AI PROFILING & OPTIMIZATION

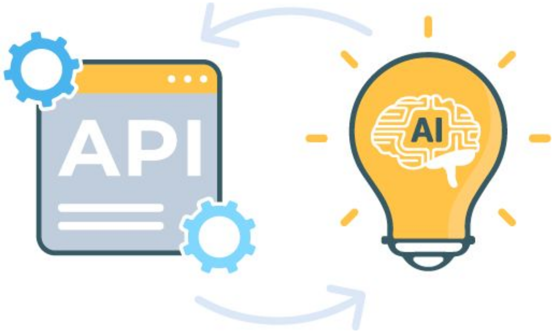
MLOps

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

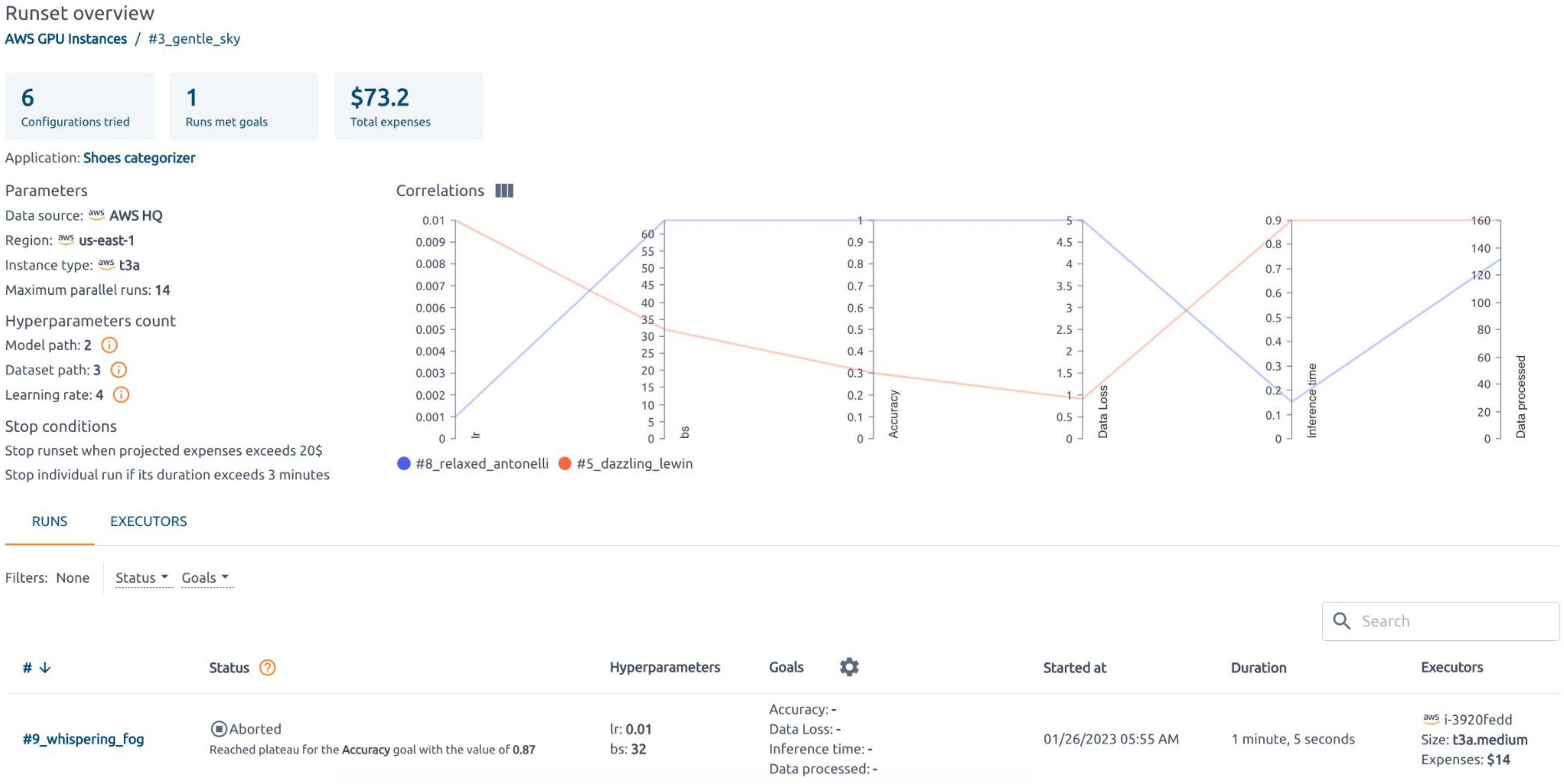


Runsets

- 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 / Saving 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

RUNS EXECUTORS

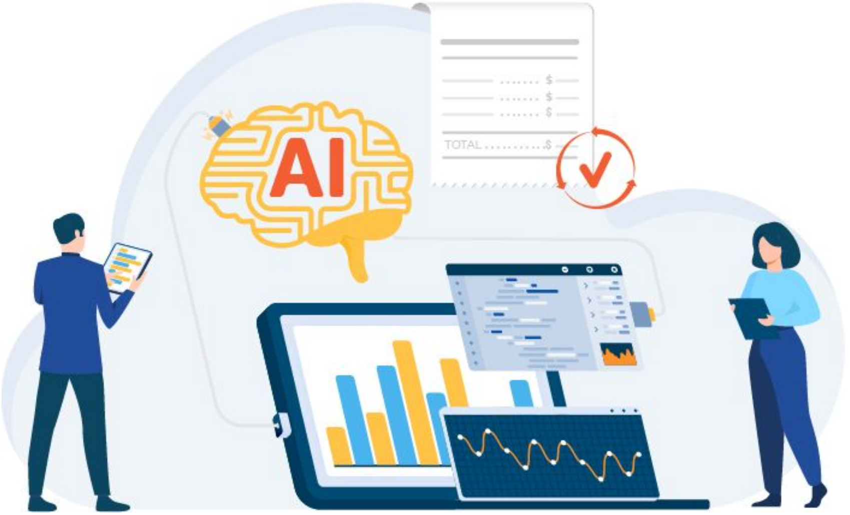
Filters: None Status ▾ Goals ▾

Search

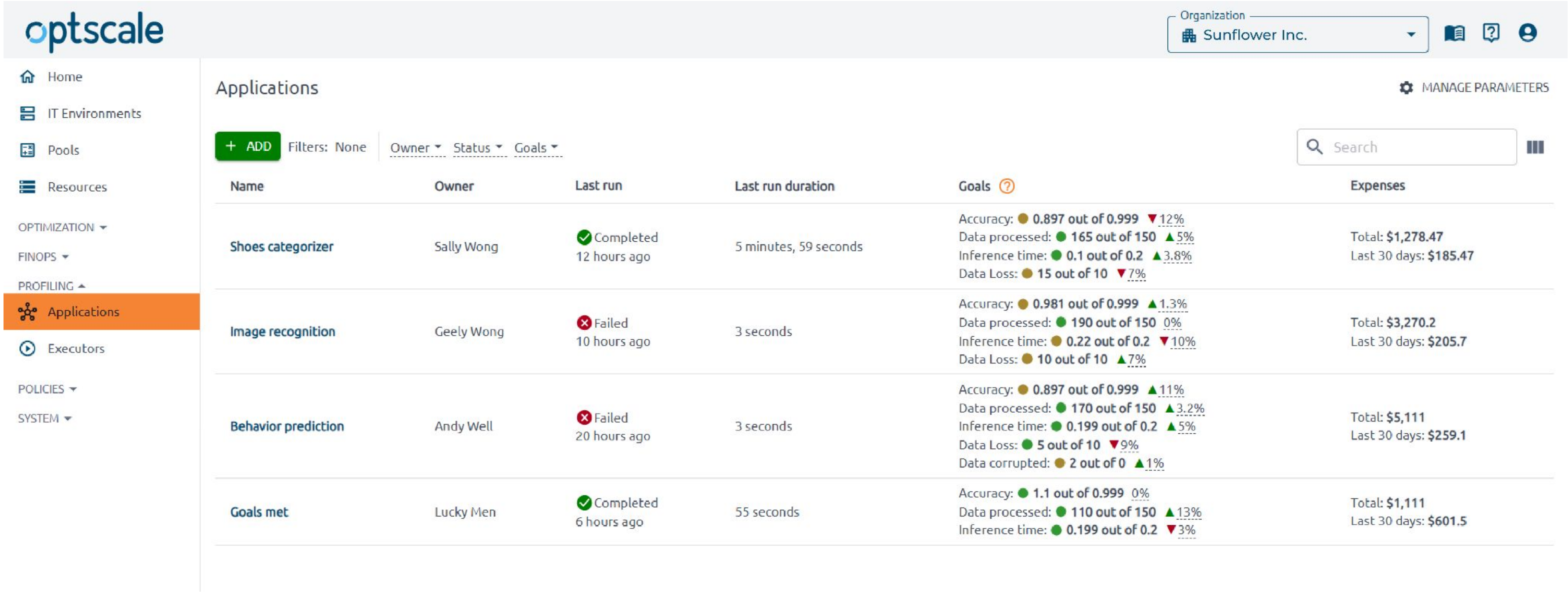
# ↓	Status ⓘ	Hyperparameters	Goals ⚙	Started at	Duration	Executors
#9_whispering_fog	Aborted Reached plateau for the **Accuracy** goal with the value of **0.87**	lr: **0.01** bs: **32**	Accuracy: - Data Loss: - Inference time: - Data processed: -	01/26/2023 05:55 AM	1 minute, 5 seconds	aws i-3920fedd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

ML R&D status observability

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



ML R&D status observability



Applications

Filters: None

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/Saving Plans and dozens of optimization scenarios

Supported platforms:

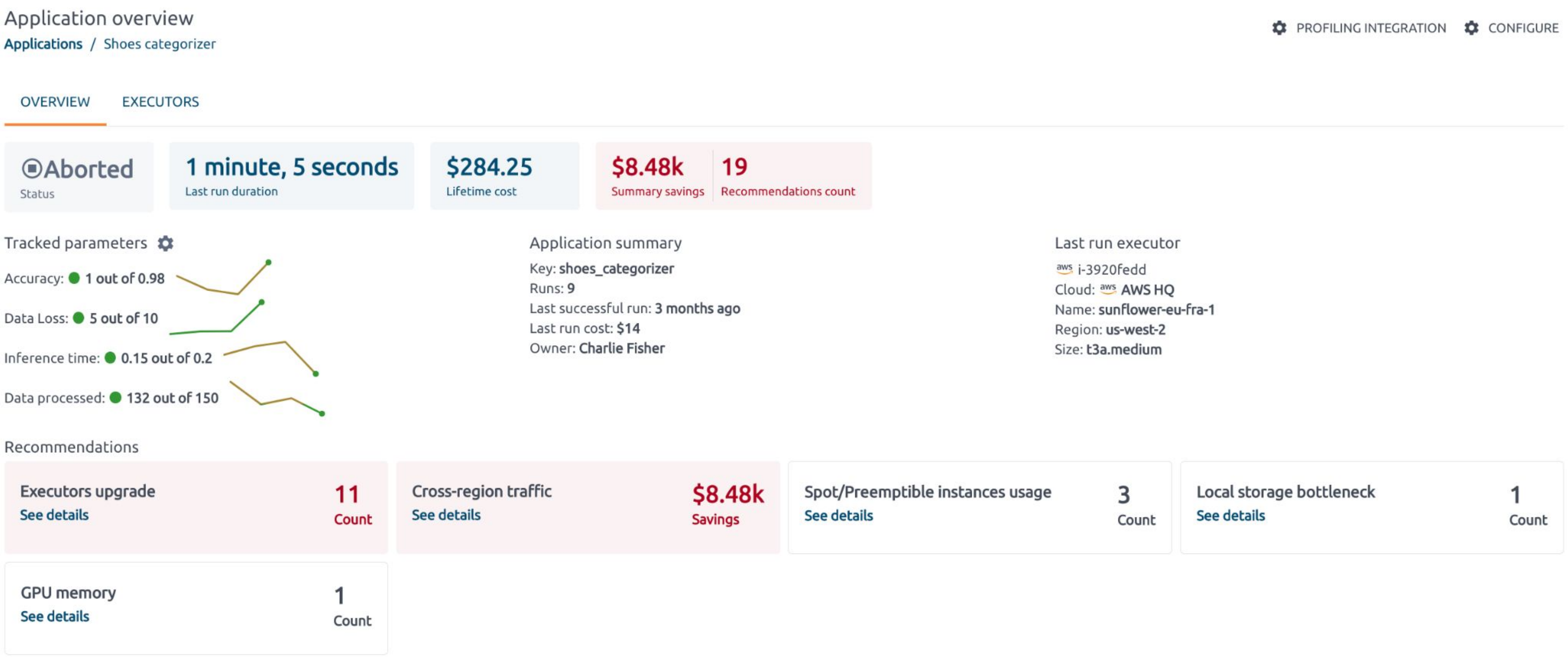
aws Microsoft Azure Google Cloud Alibaba Cloud kubernetes

databricks PyTorch Kubeflow Apache Spark TensorFlow

ML/AI optimization recommendations

- Utilizing Reserved/Spot instances and Saving Plans
- Rightsizing and instance family migration
- Detecting CPU, GPU, RAM and IO bottlenecks
- Cross-regional traffic
- Spark executor idle state
- 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

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

Recommendations

Executors upgrade
See details
11
Count

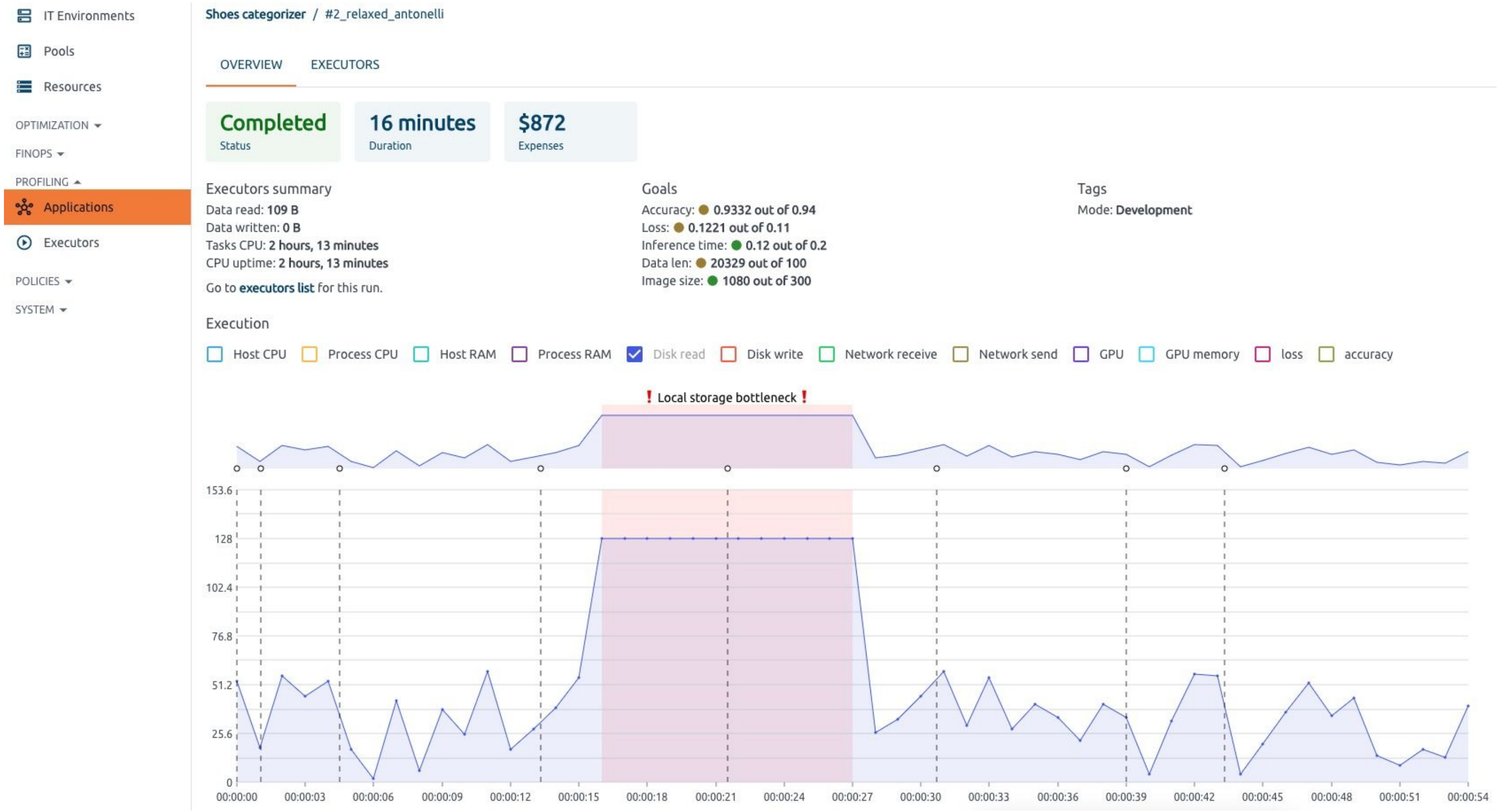
Cross-region traffic
See details
$8.48k
Savings

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



IT Environments

Pools

Resources

OPTIMIZATION

FINOPS

PROFILING

Applications

Executors

POLICIES

SYSTEM

Shoes categorizer / #2_relaxed_antonelli

OVERVIEW EXECUTORS

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 ❗

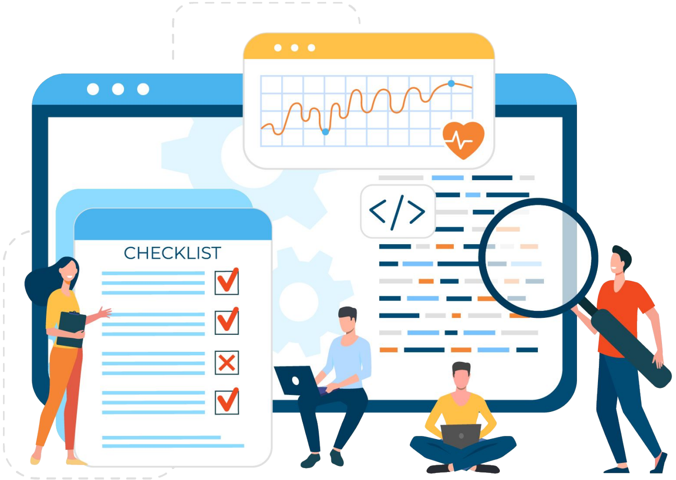
hystax

IT ENVIRONMENT MANAGEMENT

hystax

IT Environment Management

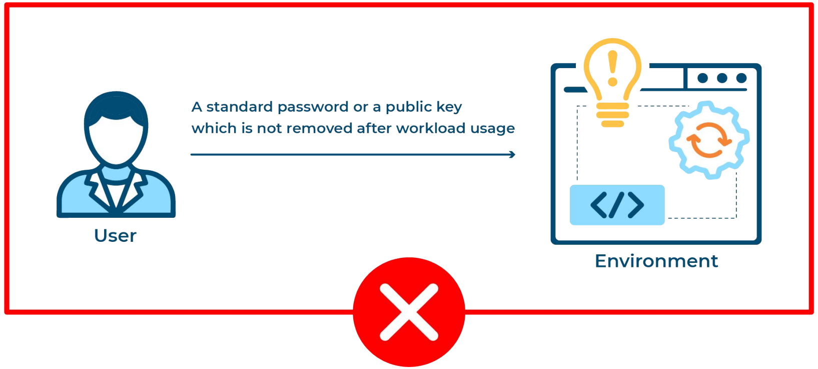
- Manage a list of IT environments, their health and availability
- Book IT environments and organize shared usage
- Track deploy history, review software versions
- Resource planning via Jira, Slack or OptScale UI
- Power management and cost optimization
- Environment performance monitoring



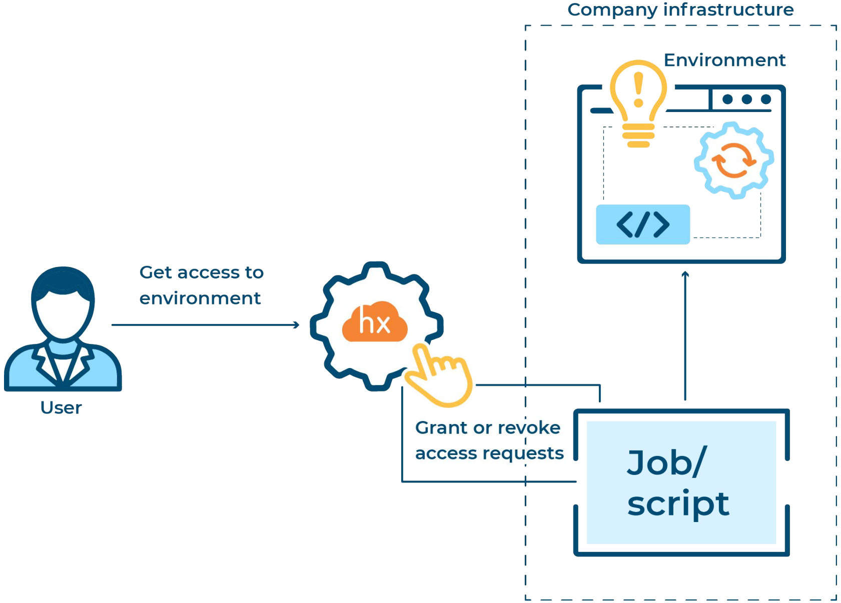
Integrations: Jira Software, slack, Jenkins, HashiCorp Terraform, GitLab, GitHub

Environment access management

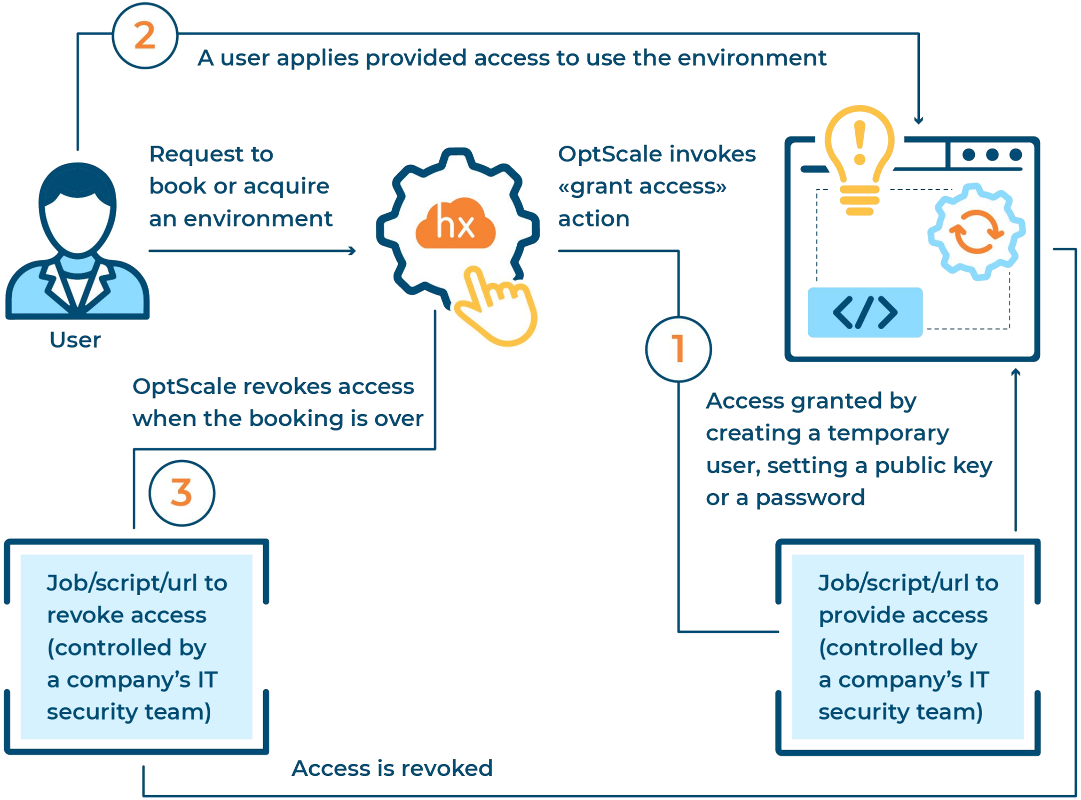
Traditional environment access management flow


User
A standard password or a public key which is not removed after workload usage
Environment

Environment access management flow with OptScale



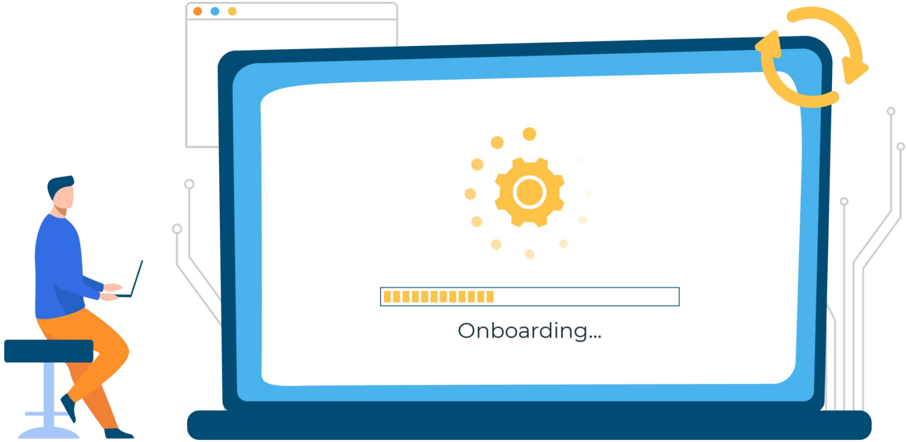
Temporary and revocable access


2
A user applies provided access to use the environment
Request to book or acquire an environment
OptScale invokes «grant access» action
User
hx
1
OptScale revokes access when the booking is over
Access granted by creating a temporary user, setting a public key or a password
3
Job/script/url to revoke access (controlled by a company's IT security team)
Job/script/url to provide access (controlled by a company's IT security team)
Access is revoked

- OptScale controls access to internal workloads
- Script or hook is invoked when a user requests access. The script to provide temporary access is owned by a company's IT security team
- When a user is done with workloads, another hook is invoked to revoke access
- Audit logs are available
- Script samples are available for a quick setup

OptScale onboarding

- **UI and API**
  UI to manage settings and view reports,
  API to integrate with jobs and pipelines
- **Ease of use. R&D tools integration**
  Your team doesn't need to learn a new tool. 90% of the functionality is available via Jira & Slack
- **SaaS or a private deployment**
  The product is available in two options
- **5 minutes to set up**
  No long configuration and deployments



hystax

Resource sharing and lifecycle management

- **Resource grouping and ownership**
  Represent clusters, stacks, jobs but not just individual resources. Acquire, release and schedule shared usage

- **TTL rules**
  TTL rules for individual resources, groups and budgets

- **Tag policies and resource auto-assignment**
  Set and manage tag rules and automatically assign resources to groups or budgets

- **Ease of use**
  Manage TTLs and other resource parameters via Slack

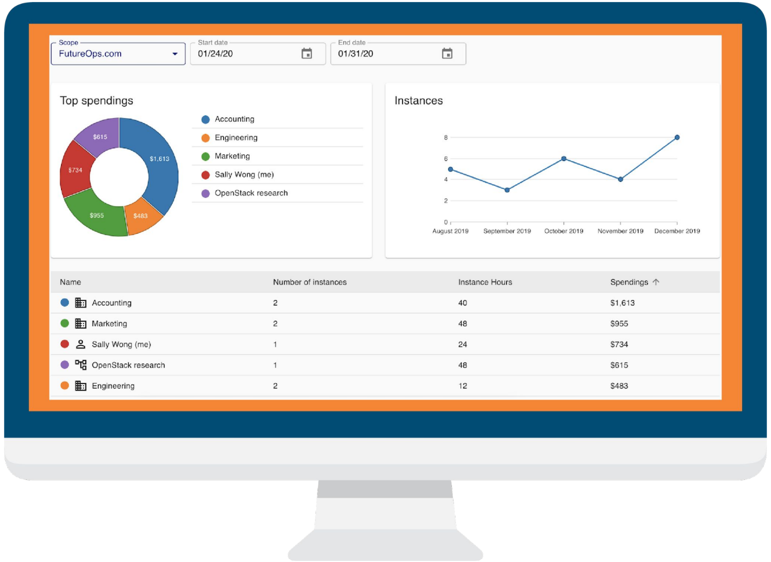
Cloud cost transparency and optimization

- **Budgets and resource auto-assignment**
  Track every BU, team, user or app expense

- **Cost anomaly detection**
  Spike identification and instant alerts

- **In-depth cost analysis**
  Budget and resource expense history

- **Budget forecasts, thresholds & insights**
  Get analytics, forecast and optimization insights



Scope: FutureOps.com
Start date: 01/24/20
End date: 01/31/20

Name	Number of instances	Instance Hours	Spendings ↑
● Accounting	2	40	$1,613
● Marketing	2	48	$955
● Sally Wong (me)	1	24	$734
● OpenStack research	1	48	$615
● Engineering	2	12	$483

hystax

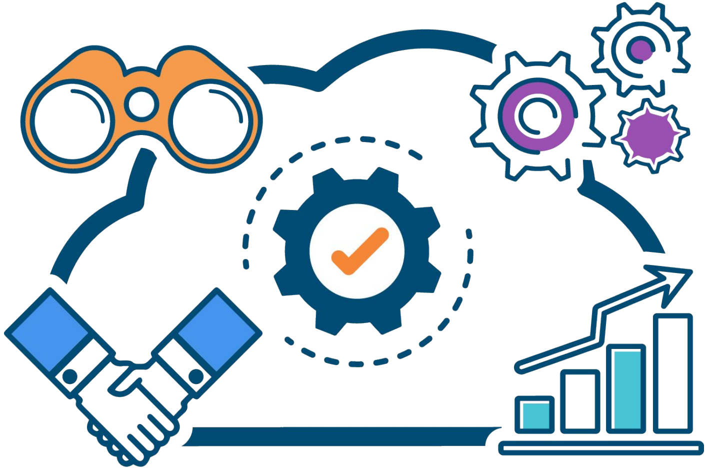
FinOps enablement

- **Product to build a FinOps process**
  Visibility, Optimization, Control and Collaboration

- **Engineering engagement**
  Team members are responsible for their resources, TTLs and cloud spending

- **No new tools onboarding.**
  Just collaborate via Slack
  Destroy, notify, notify & destroy scenarios

- **OptScale conforms with FinOps practices**
  Hystax leads one of the biggest FinOps communities



Contacts

☎ +1 628 251 1280

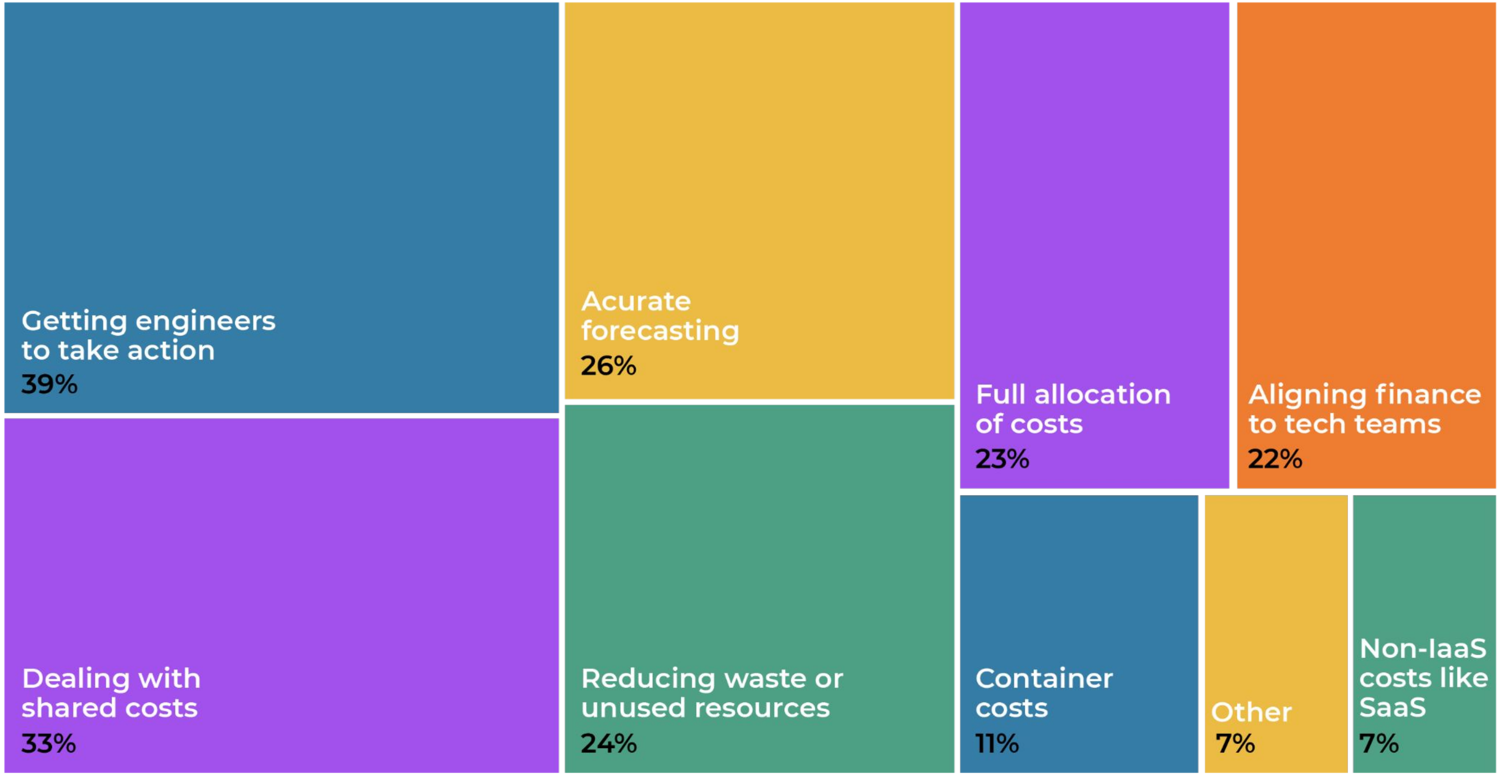
🌐 hystax.com

✉ info@hystax.com

📍 1250 Borregas Avenue, Sunnyvale, CA, 94089

BACKUP SLIDES

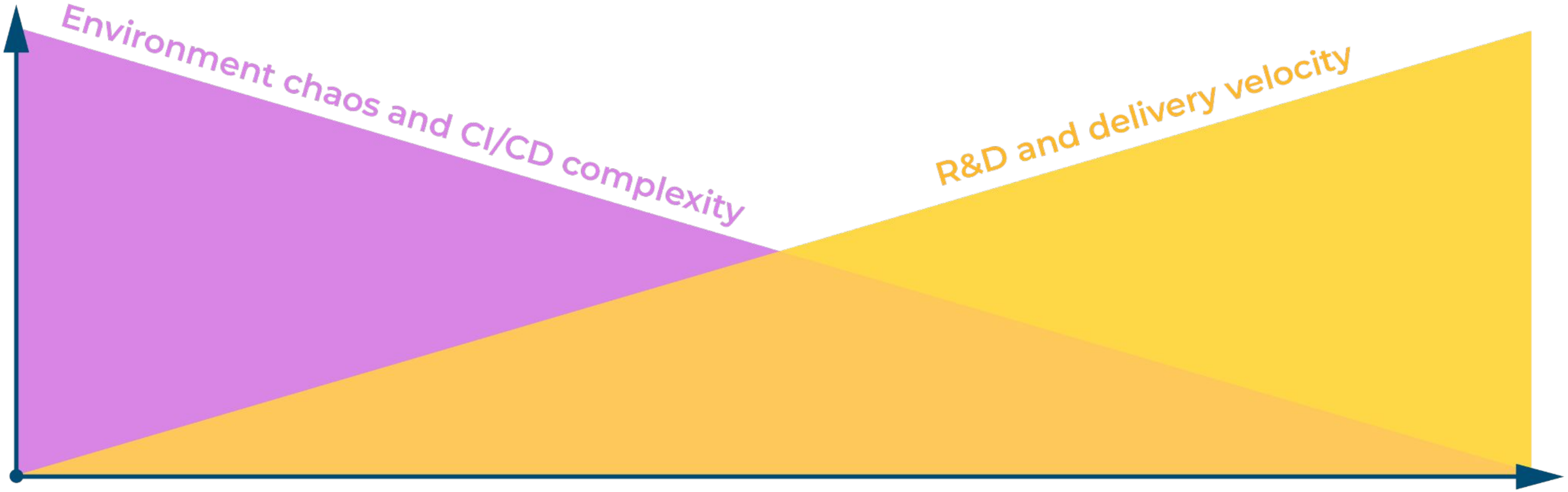
FinOps adoption challenges



The State of FinOps Report 2021 | FinOps Foundation | www.finops.org

hystax

Test Environment Management



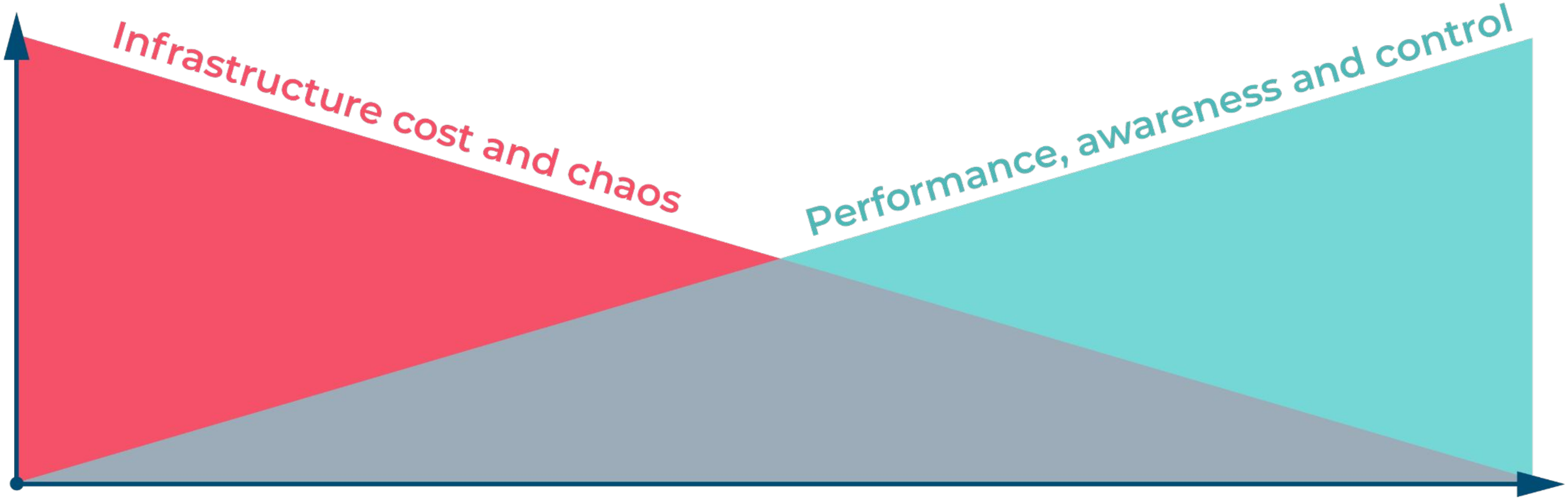
Environment visibility, booking and shared usage optimization

Software version tracking and environment performance monitoring

Integration with Jira, Jenkins and Slack

Accurate cost of delivery forecast

FinOps


Infrastructure cost and chaos
Performance, awareness and control

Cost management and optimization

Cost allocation and accurate forecasts

Engineering teams engagement

Cloud cost management issues. Why FinOps?

- **Engineers are not engaged in cost-saving processes**

  Getting a long list of optimization scenarios does not help as a few DevOps or Central IT folks cannot fix all the optimization issues without reaching resource owners who have other priorities.
  As a result, only 20-30% 'low-hanging fruit' recommendations are implemented

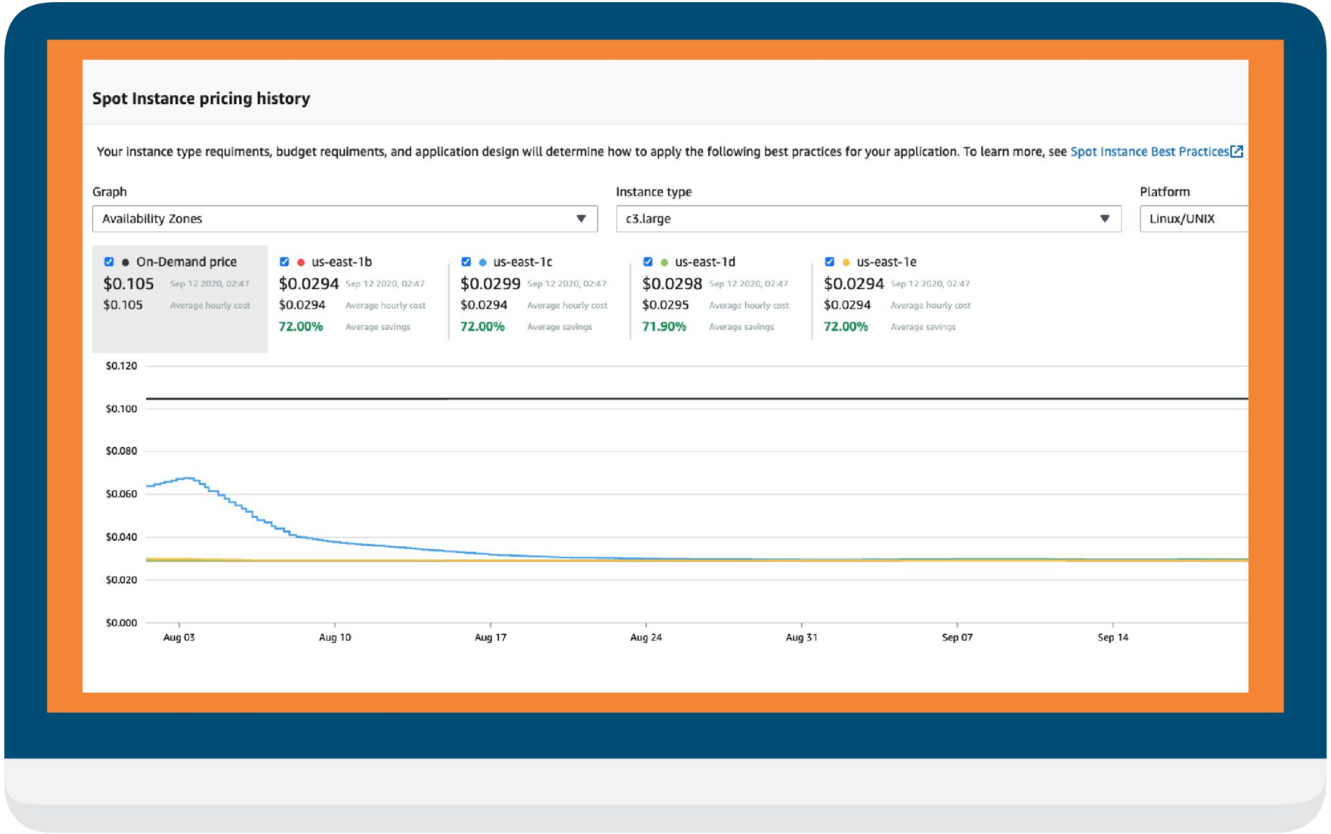
- **No resource lifecycle management**

  Cloud cost management tools don't give a way to manage resource lifecycle management

- **No transparency and flexibility**

  Cloud-native tools do not provide enough granularity and transparency across budgets, teams, clusters and applications

Spot Instance Price Variation



Spot Instance pricing history

Your instance type requiments, budget requiments, and application design will determine how to apply the following best practices for your application. To learn more, see Spot Instance Best Practices

Graph: Availability Zones ▼

Instance type: c3.large ▼

Platform: Linux/UNIX

☑ ● On-Demand price	☑ ● us-east-1b	☑ ● us-east-1c	☑ ● us-east-1d	☑ ● us-east-1e
$0.105 Sep 12 2020, 02:47	$0.0294 Sep 12 2020, 02:47	$0.0299 Sep 12 2020, 02:47	$0.0298 Sep 12 2020, 02:47	$0.0294 Sep 12 2020, 02:47
$0.105 Average hourly cost	$0.0294 Average hourly cost	$0.0294 Average hourly cost	$0.0295 Average hourly cost	$0.0294 Average hourly cost
	72.00% Average savings	72.00% Average savings	71.90% Average savings	72.00% Average savings

What’s new (May 2023)

- Reserved Instances/Saving Plans visualization
- Anomaly detection and constraints
- ML or any application profiling & optimization
- MLOps
- Scalability and performance improvements

Roadmap (open source and SaaS versions)

- RI/SP/Spot recommendation improvement
- Rightsizing with RAM and GPU
- S3 duplicates, tiering, profiling