

Хайстекс Оptsкейл

FinOps и MLOps платформа с
открытым исходным кодом

Запускайте ML/AI или любой тип рабочей
нагрузки с оптимальной производительностью
и стоимостью инфраструктуры



Хайстекс



Компания основана в 2016 году,
разработка и поддержка в РФ



Среди наших клиентов -
Сбербанк, Яндекс.Облако, X5
Retail Group, Burger King, М.
Видео, Спортмастер

Сценарии использования ОптСкейл

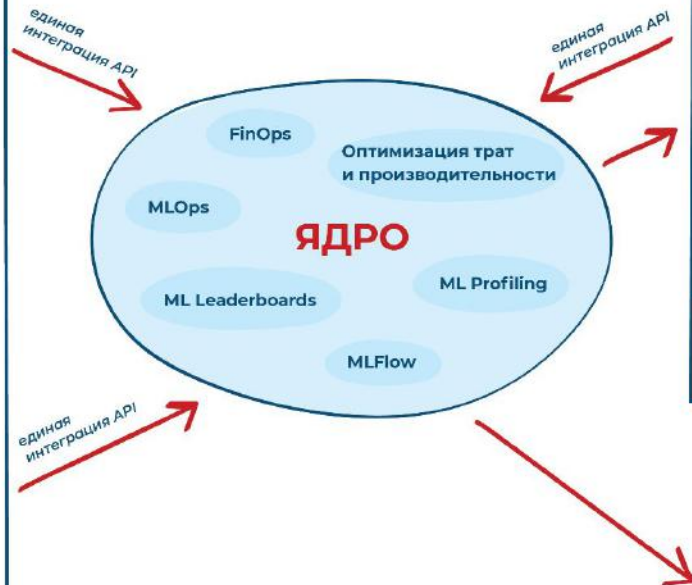


Источники данных



* - в процессе разработки

Рабочая схема ОптСкейл



Рекомендации

- S3 дедупликация
- VM Rightsizing
- Неиспользуемые ресурсы
- RDS, SageMaker instance family selection and rightsizing
- Reserved Instance/ Saving Plans and Spot Instance usage
- Power management
- Рекомендации по безопасности

Любые кастомные рекомендации

Дашборды

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

FinOps и оптимизация расходов на облако

FinOps и оптимизация расходов на облако

- Прогнозирование и мониторинг затрат на ИТ-инфраструктуру
- Выявление потерь и оптимизация расходов
- Обеспечение прозрачности ресурсов/приложений/сервисов
- Управление ИТ-ресурсами
- Установка TTL и ограничений бюджета
- Внедрение долгосрочного процесса FinOps путем вовлечения команд инженеров

Хайстекс ОптСкейл vs cloud-native решения

- Прозрачность облачных ресурсов, включая различные облака, учетные записи и регионы
- Десятки сценариев оптимизации, не поддерживаемых облачными провайдерами, включая один из лучших механизмов rightsizing (перераспределение ресурсов инфраструктуры в зависимости от текущих потребностей)
- Распределение затрат не только по тегам, но и по другим свойствам
- Карта географического распределения и сетевого трафика
- Правила TTL и ограничений бюджета
- FinOps: ОптСкейл создан для инженеров, ответственных за свои облачные ресурсы

Оптимизация расходов на облако vs FinOps

Оптимизация расходов:

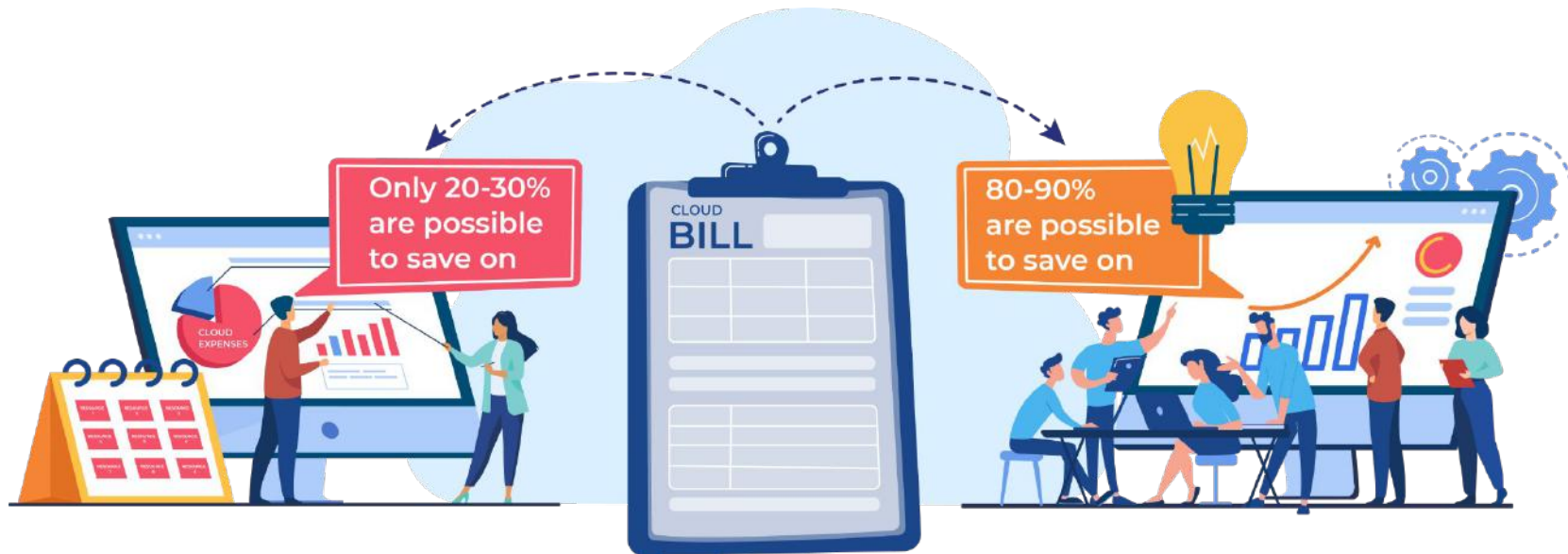
- Фокус на ИТ сотруднике, которому необходимо отслеживать R&D команды для тегирования, перераспределения, удаления неиспользуемых ресурсов
- Предоставление отчета, для решения проблем в краткосрочной перспективе, через несколько месяцев проблемы возвращаются
- R&D команда отключена от процесса экономии затрат и не несет ответственности

FinOps:

- Фокус на всей команде FinOps, включая инженеров, которые генерируют основную часть затрат
- Создание долгосрочного процесса экономии расходов путем вовлечения и обучения команды
- ИТ-специалисты несут ответственность за процесс управления; инженеры - за свои ресурсы и TTL; ОптСкейл - за обучение команд и предоставление эффективных практик

Оптимизация расходов на облако

FinOps



Облачные решения по управлению затратами созданы только для нескольких ИТ-сотрудников, ответственных за оптимизацию затрат, но у них ограничены управленческие функции и влияние на команды исследования и разработки

FinOps вовлекает в процессы оптимизации затрат руководителей компаний, финансовые и инженерные команды

MLOPS: ML/AI профайлинг и оптимизация

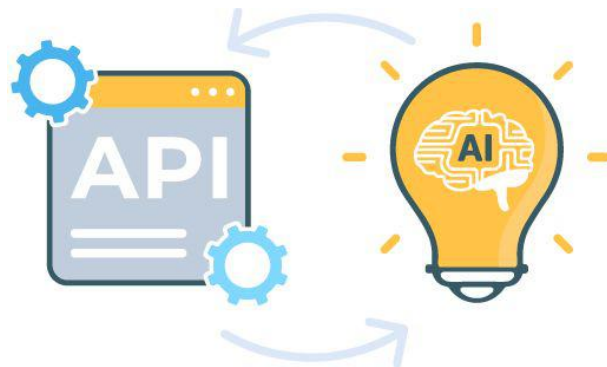
MLOps

- Запуск Runsets для автоматического масштабирования количества экспериментов
- Наблюдение за прогрессом команды и индивидуальных инженеров в области машинного обучения
- Профилирование задач ML/AI, выявление узких мест
- Рекомендации по оптимизации



Runsets

- Автоматический запуск нескольких экспериментов с настраиваемыми наборами данных, диапазонами гиперпараметров и версиями моделей
- Оптимальное использование оборудования с эффективным использованием Spot, Reserved Instances/Saving Plans
- Настроенные цели экспериментов и критерии успеха
- Различные условия завершения/прерывания - выбор первого успешного, завершение всех
- Интегрированное профилирование для выявления узких мест



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: **us-east-1**

Instance type: **t3a**

Maximum parallel runs: **14**

Hyperparameters count

Model path: **2**

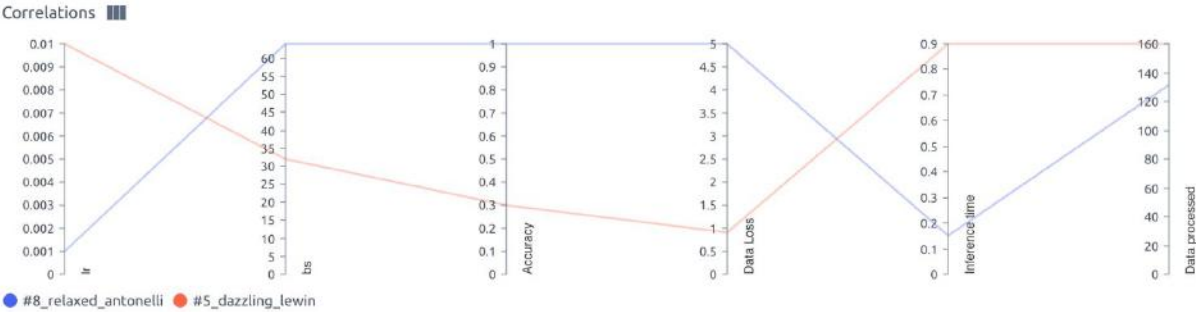
Dataset path: **3**

Learning rate: **4**

Stop conditions

Stop runset when projected expenses exceeds 20\$

Stop individual run if its duration exceeds 3 minutes



RUNS EXECUTORS

Filters: None Status Goals

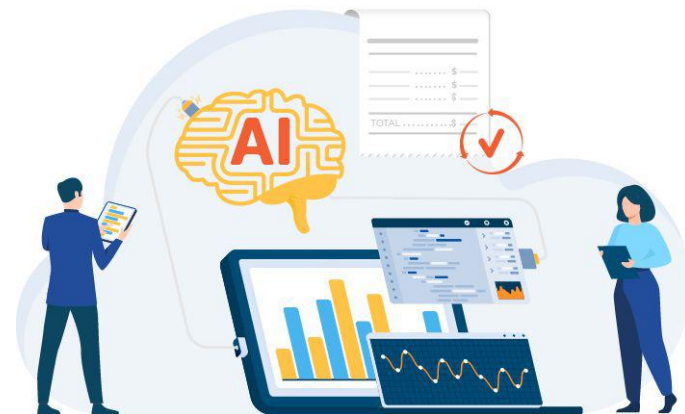
Q

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

Прозрачность состояния ML R&D целей

- Список моделей с состоянием целей и активными рекомендациями
- Отслеживание количества и качества проведенных командой экспериментов
- Стоимость общей модели и отдельных экспериментов



Прозрачность состояния ML R&D целей

optscale

Organization

Sunflower Inc.

📖

?

👤

🏠 Home

📁 IT Environments

📁 Pools

📁 Resources

OPTIMIZATION ▾

FINOPS ▾

PROFILING ▲

🔗 Applications

🕒 Executors

POLICIES ▾

SYSTEM ▾

Applications

+ ADD

Filters: None

Owner ▾

Status ▾

Goals ▾

🔍 Search

☰

MANAGE PARAMETERS

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 профайлинг и оптимизация

- Отслеживание и профилирование обучения моделей ML/AI, сбор метрик внутри и снаружи
- Отслеживание корреляции CPU/RAM/GPU/Disk IO
- Минимальные затраты на облачные ресурсы для экспериментов и разработки ML/AI за счет использования Reserved Instances/Saving Plans и десятков других сценариев оптимизации

ML/AI оптимизационные рекомендации

- Использование Reserved/Spot instances и Saving Plans
- Rightsizing и миграция инстансов
- Выявление узких мест в CPU, GPU, RAM и IO
- Межрегиональный трафик
- Сравнение экспериментов/запусков



ML/AI профайлинг и оптимизация

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 профайлинг и оптимизация

IT Environments

Pools

Resources

OPTIMIZATION ▾

FINOPS ▾

PROFILING ▴

Applications

Executors

POLICIES ▾

SYSTEM ▾

Shoes categorizer / #2_relaxed_antonnelli

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

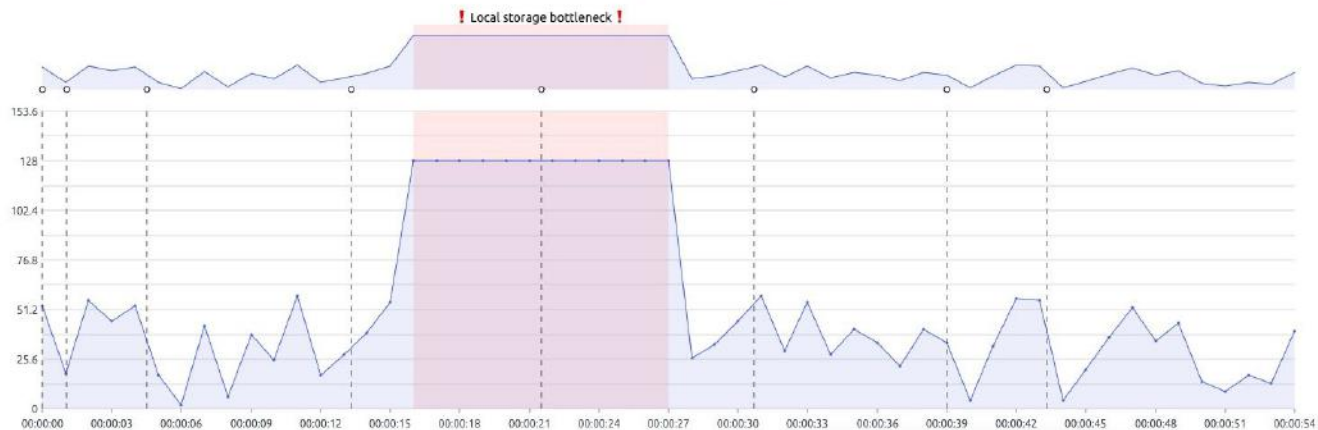
Image size: ● 1080 out of 300

Tags

Mode: Development

Execution

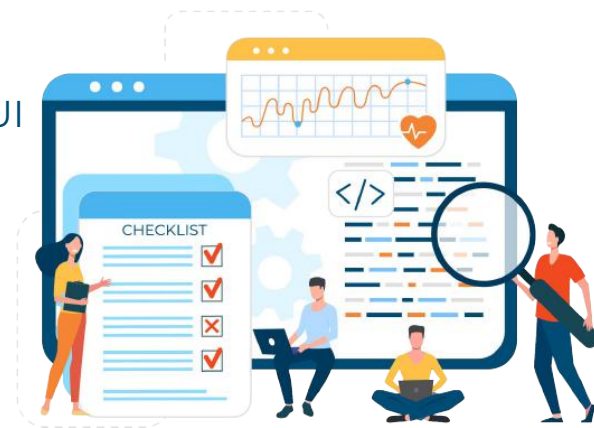
Host CPU Process CPU Host RAM Process RAM Disk read Disk write Network receive Network send GPU GPU memory loss accuracy



Управление ИТ-окружениями

Управление ИТ-окружениями

- Управление списком ИТ-окружений, их состоянием и доступностью
- Бронирование ИТ-окружений и организация общего использования
- Отслеживание истории развертывания, просмотр версий программного обеспечения
- Планирование ресурсов через Jira, Slack или ОптСкейл UI
- Эффективное управление и оптимизация затрат
- Мониторинг производительности ИТ-окружений

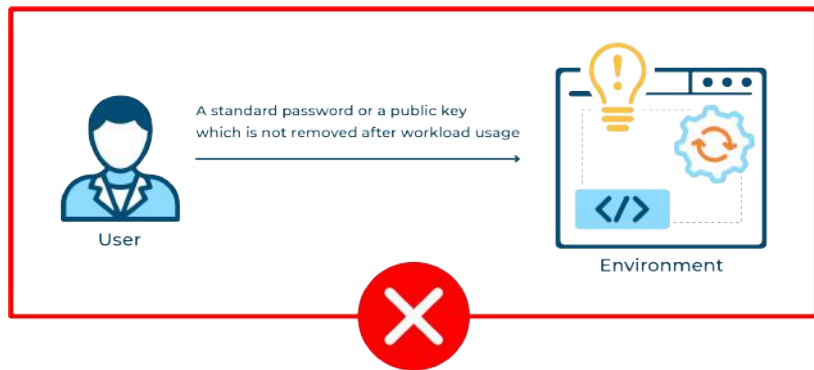


Интеграции:  Jira Software

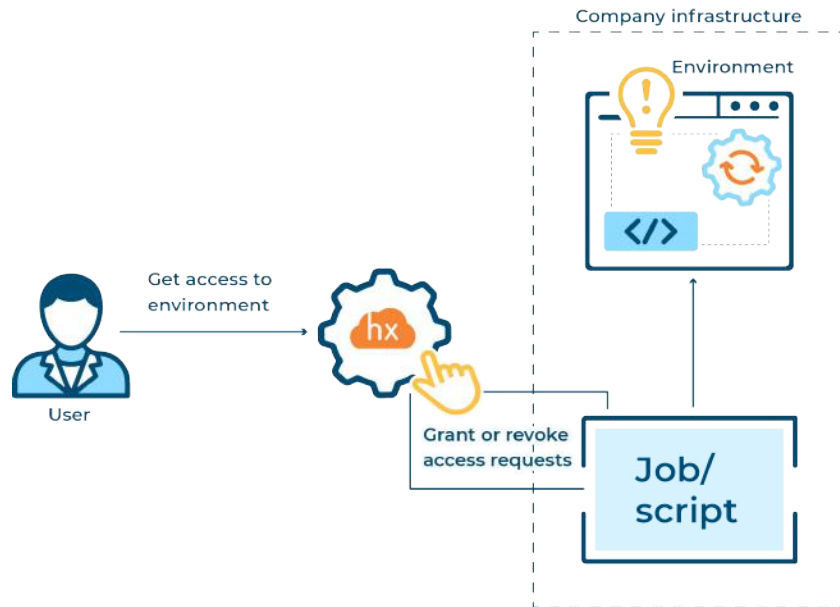


Управление доступом к ИТ-окружениям

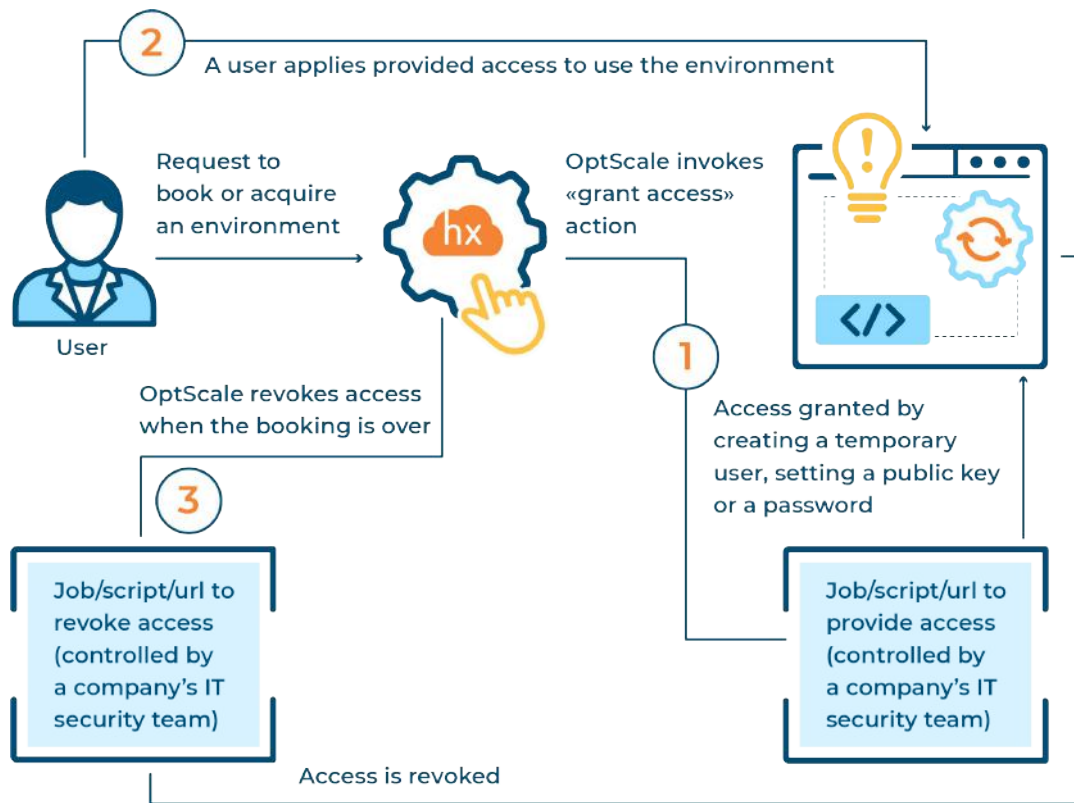
Традиционная система управления доступом к ИТ-окружениям



Система управления доступом к ИТ-окружениям с OptScale



Временный и отзываемый доступ



- ОптСкейл контролирует доступ к внутренним рабочим нагрузкам
- Скрипт или hook вызывается, когда пользователь запрашивает доступ. Сценарий предоставления временного доступа принадлежит самой компании
- Когда пользователь завершает работу с нагрузками, вызывается другой hook для отзыва доступа
- Доступны журналы аудита
- Доступны образцы скриптов для быстрой настройки

Внедрение ОптСкейл

- **UI и API**

UI для управления настройками и просмотра отчетов,
API для интеграции с джобами и пайплайнами

- **Простота в использовании. Интеграция с R&D инструментами**

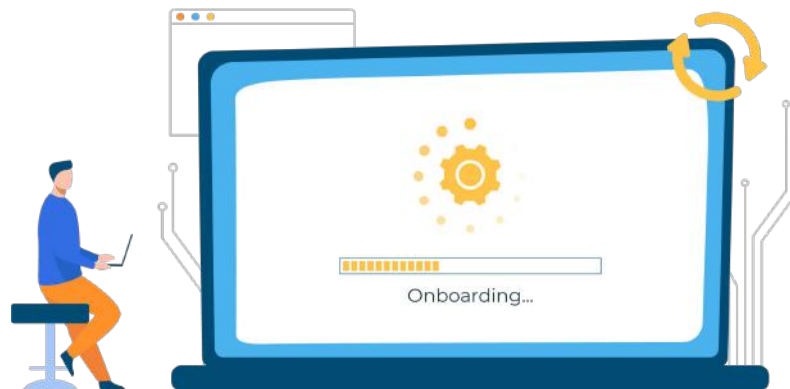
Вашей команде не нужно изучать новый инструмент.
90% функционала доступно через Jira и Slack

- **SaaS или частное развертывание**

Две опции доступны в продукте

- **5 минут для настройки**

Нет долгой конфигурации и развертывания



Совместное использование ресурсов и управление их жизненным циклом

- **Группировка ресурсов и установление их владельцев**

Представление кластеров, стеков, задач, а не только отдельных ресурсов.
Получение, освобождение и планирование общего использования

- **Правила TTL**

Правила TTL для отдельных ресурсов, групп и бюджетов

- **Политики тегов и автоматическое назначение ресурсов**

Установление и управление правилами тегов и автоматическое назначение ресурсов группам или бюджетам

- **Простота использования**

Управление TTL и другими параметрами ресурсов через Slack



Прозрачность и оптимизация облачных затрат

- **Бюджеты и автоматическое назначение ресурсов**

Отслеживание затрат каждого бизнес-юнита, команды, пользователя или приложения

- **Обнаружение аномалий в затратах**

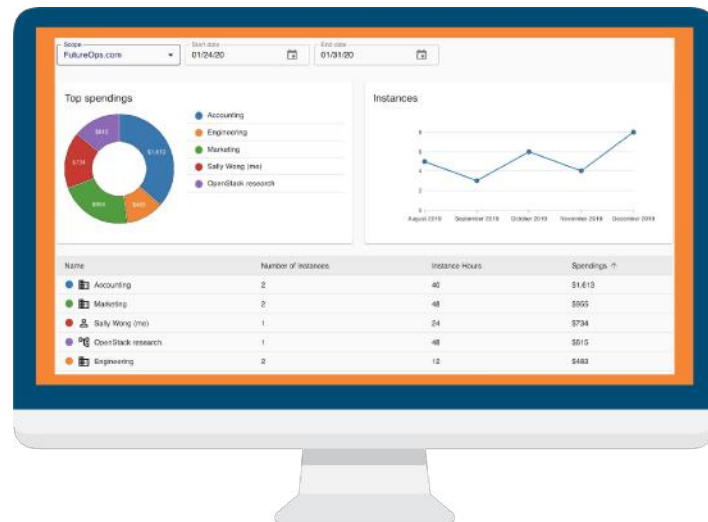
Выявление пиков и мгновенные оповещения

- **Глубокий анализ затрат**

История расходов и бюджетов на ресурсы

- **Прогнозы бюджетов, пороговые значения и аналитика**

Получение аналитики, прогнозов и рекомендаций по оптимизации



Внедрение FinOps процесса

- **Продукт для построения FinOps**
Видимость, оптимизация, контроль и сотрудничество
- **Вовлечение инженеров**
Члены команды несут ответственность за свои ресурсы, TTL и расходы в облаке
- **Нет необходимости в внедрении новых инструментов. Используйте Slack**
Сценарии Уничтожения, Уведомления, Уведомить и Уничтожить



Текущие клиенты и партнеры

Telecoms



Integrators



Cloud service providers



Cloud vendors



End customers



Контакты

✉ sales@hystax-team.ru



www.xct.pф

☎ +7 495 204 28 77



123112, Москва, Пресненская
наб., д.12, эт/пом/ком/45/1, 2/82

Приложение

Сложности в управлении затратами на облако. Почему FinOps?

- **Инженеры не занимаются процессами оптимизации затрат.**

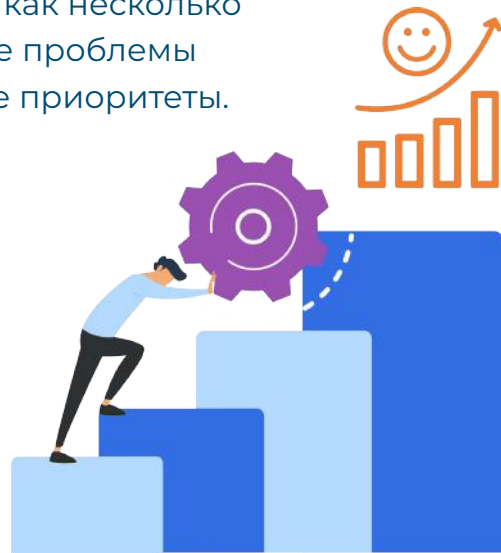
Получение длинного списка сценариев оптимизации не помогает, так как несколько сотрудников DevOps или руководители ИТ-команд не могут решить все проблемы оптимизации без связи с владельцами ресурсов, у которых есть другие приоритеты. В результате реализуется только 20-30% рекомендаций.

- **Отсутствие управления жизненным циклом ресурсов**

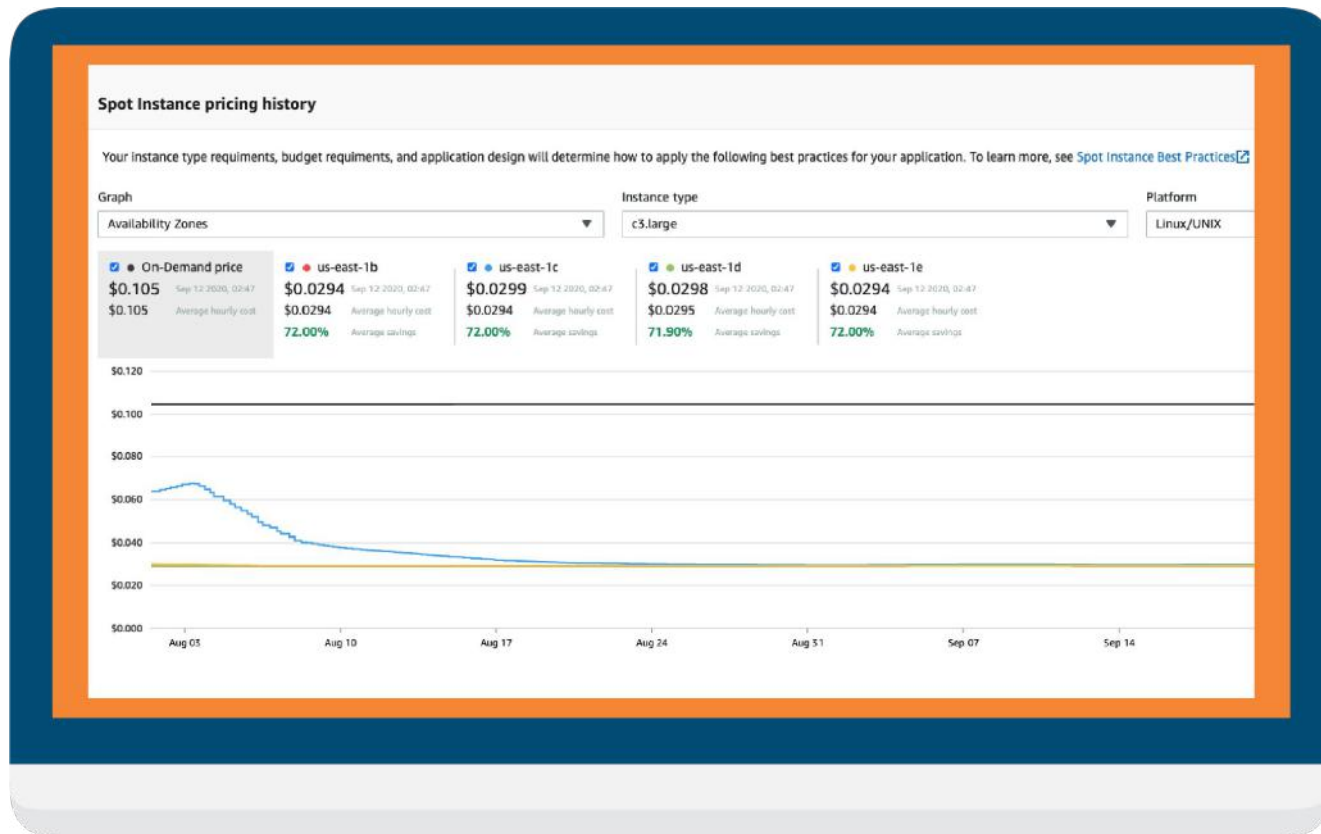
Инструменты управления затратами в облаке не предоставляют возможности управления жизненным циклом ресурсов.

- **Отсутствие прозрачности и гибкости**

Инструменты, разработанные облачными провайдерами, не обеспечивают достаточной детализации и прозрачности в рамках бюджетов, команд, кластеров и приложений.



Spot Instance - диапазон цен



Что нового? (Май 2023)

- Прозрачность и визуализация Reserved Instances/Saving Plans
- Обнаружение аномалий и ограничений
- Профилирование и оптимизация ML или любых других приложений
- MLOps
- Масштабируемость и повышение производительности

Roadmap (версии с открытым исходным кодом и SaaS)

- Улучшение рекомендаций по RI/SP/Spot
- Оптимизация с учетом RAM и GPU
- S3 дубликаты, тиринг, профайлинг

