Идеальная «песочница» для ML-моделей:

настраиваем контейнеризацию без стресса

Даниил Салман

Tech Lead по практике контейнеризации
K2 Tech

K2TEX

K2TEX

Эксперты цифровизации бизнеса

18
лет на рынке

1 500+
ИТ-профессионалов

250+
ведущих инженеров, senior-разработчиков и экспертов

1 200+
проектов в год в сфере ИТ-услуг

10+
лет — средний стаж работы специалистов

9
собственных ИТ-продуктов

Помогаем клиентам достигать бизнес-целей, используя ИТ-инструменты. Внедряем технологические решения российских и зарубежных производителей, а также разрабатываем специализированные цифровые продукты.

Сегодня разберём:

С какими болями сталкиваются ML-команды каждый день

Почему «просто k8s-кластер» не рабочее решение

Какой стек нужен, чтобы собрать крепкий скелет ML-платформы

Практика: от запуска обучения через Airflow до деплоя и инференса модели в k8s

Сегодня разберём:

Кому будет интересно:

ML-инженеры, Data Scientists, архитекторы
и инфраструктурные команды

Что стоит знать заранее:

Базовое понимание Kubernetes, Docker
и Python-фреймворков

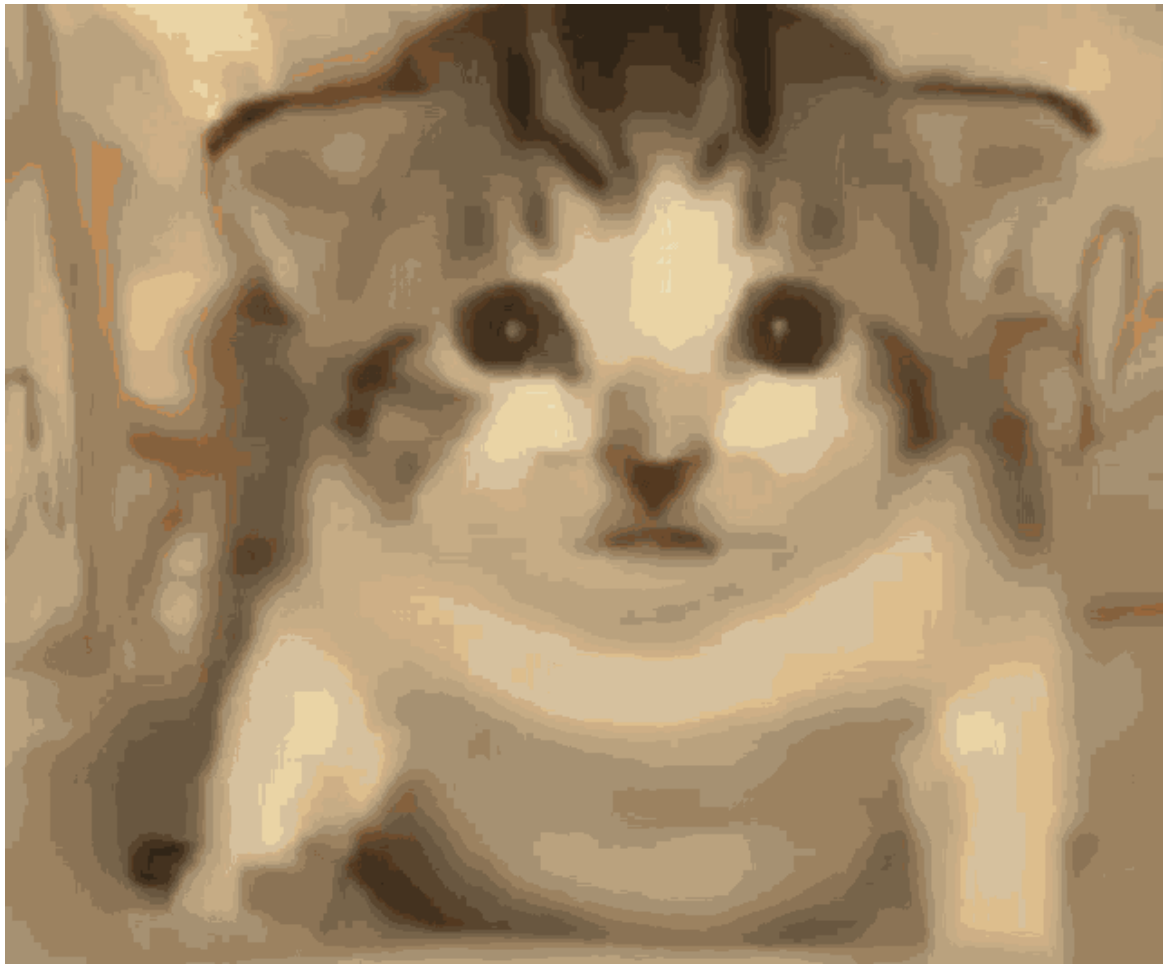
Что вы получите к концу доклада:

- Разберётесь, как собрать свою ML-песочницу без боли с драйверами и CUDA
- Прогоните DAG в Airflow для обучения модели
- Научитесь централизованно логировать метрики и артефакты моделей
- Увидите, как модель превращается в inference-сервис

Немного истории Германа...

source: Sora

Герман:



source: https://is.gd/sDxhKK

Немного истории Германа...

source: Sora

source: Российское скетч-шоу «ДаЁшь молодЁжь!»

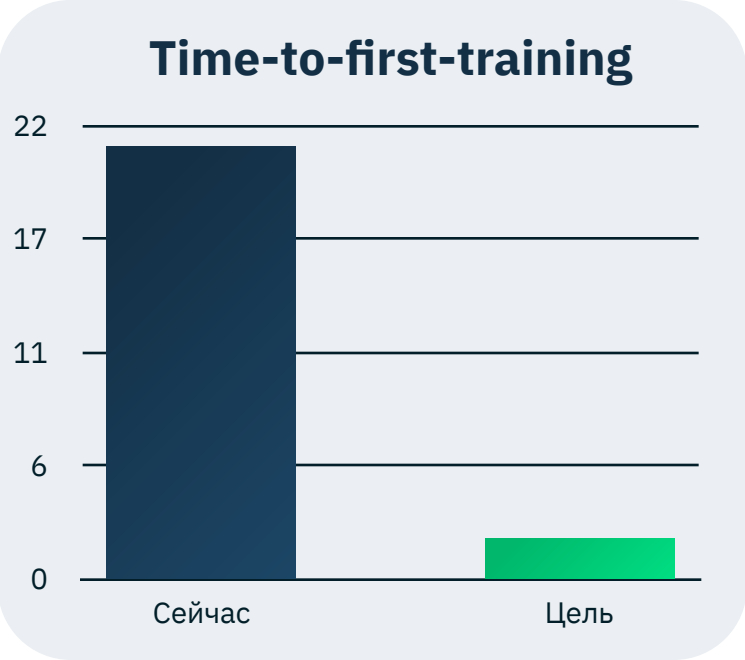
С чем сталкиваются ML-инженеры

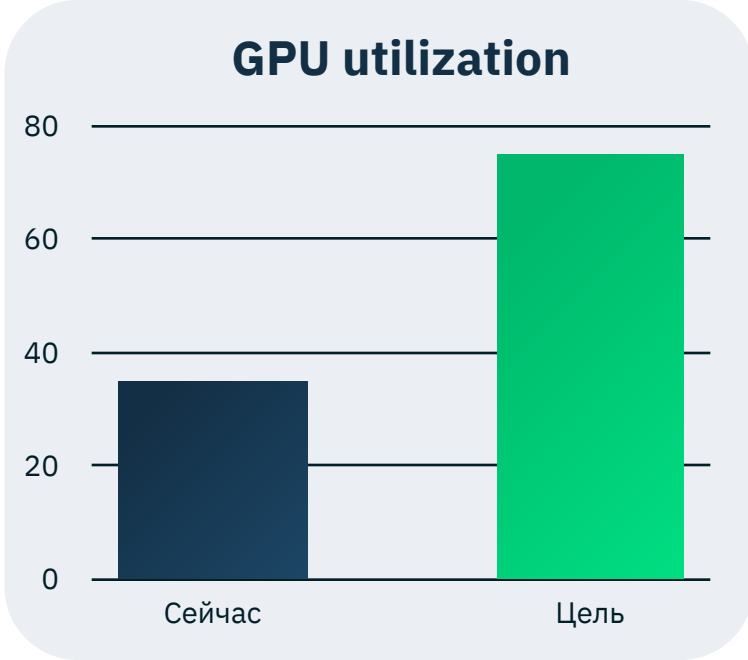
⏱️ Time-to-first-training: 2–4 недели вместо 1–2 дней

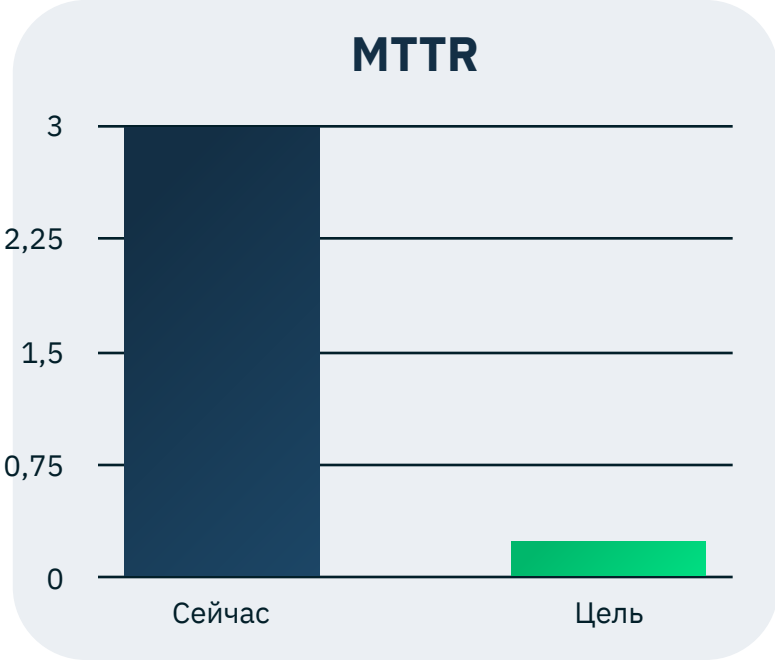
🎮 GPU utilization: 30–40% вместо 70–80%

🔄 MTTR: дни вместо часов

⚠ Каждая 3-я модель не воспроизводится в проде







Подходы к развертыванию k8s

Способ	Для кого	Затраты	Контроль	Поддержка	Мониторинг	Плюсы	Минусы
kubeadm	Энтузиасты DevOps-гуру	Низкие	Максимальный	Сам	Свой	Максимальная гибкость, понимание изнутри	Много ручной работы
kuberspray	Рядовой DevOps-инженер	Средние	Высокий	Сообщество	Свой	Баланс гибкости и автоматизации, open source	Ansible
KaaS в облаке	Бизнес, готовый платить	Средние/ высокие	Средний	Облако	Облако	Удобно, быстро, поддержка	Нет всех сервисов “из коробки”, привязка к облаку
Вендорские решения	Бизнес, готовый платить	Высокие	Средний	Вендор	Встроенный	Интеграции, поддержка, все сервисы “с ходу”, быстрый старт	Стоимость, закрытость решений

Technology stack kubeadm

Observability	Security	Networking	Storage	Container Runtime
	RBAC PKI			
Service Discovery & Coordination	**Service Proxy**	**Service Mesh**	**Backup**	**OS**
CoreDNS etcd	kube-proxy			Any OS
Conf. Management	**Infrastructure providers**	**UI**	**K8S services**	**Scaling**
			kube-apiserver controller-manager scheduler etcd kubelet	

Technology stack kubespray

Observability	Security	Networking	Storage	Container Runtime
	RBAC PKI	• **Calico** • **Cilium** • **Flannel**		• **Containerd (default)** • **Cri-o** • **Docker**
Service Discovery & Coordination	**Service Proxy**	**Service Mesh**	**Backup**	**OS**
CoreDNS etcd	kube-proxy			• **Из коробки нет поддержки российских ОС**
Conf. Management	**Infrastructure providers**	**UI**	**K8S services**	**Scaling**
			kube-apiserver controller-manager scheduler etcd kubelet	

Technology stack NOVA

Observability	Security	Networking	Storage	Container Runtime
• **Prometheus** • **Grafana** • **Alertmanager** • **Thanos Query** • **Logging Operator** • **Opensearch** • **Fluentd**	• **RBAC** • **Neuvector** • **Secrets Webhook** • **CertManager** • **Secrets Manager (CSI, PKI, OAuth)**	Calico Cilium	• **Longhorn**	Containerd
Service Discovery & Coordination	**Service Proxy**	**Service Mesh**	**Backup**	**OS**
CoreDNS etcd	kube-proxy		• **Velero**	• **RED OS** • **AlmaLinux** • **Astra Linux** • **RHEL-based**
Conf. Management	**Infrastructure providers**	**UI**	**K8S services**	**Scaling**
• **Gitea** • **FluxCD** • **Authomation tools**	• **Bare Metal** • **oVirt** • **zVirt** • **VMware vSphere**	• **Nova Console**	kube-apiserver controller-manager scheduler etcd kubelet	• **VPA** • **HPA**

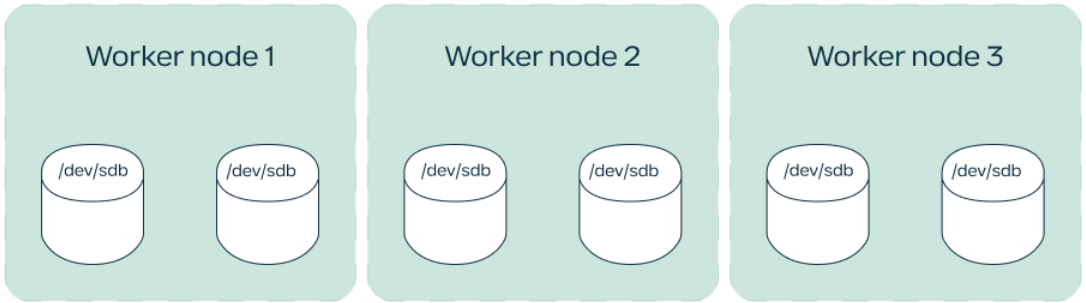
source: Sora

А что же есть в нашей песочнице?

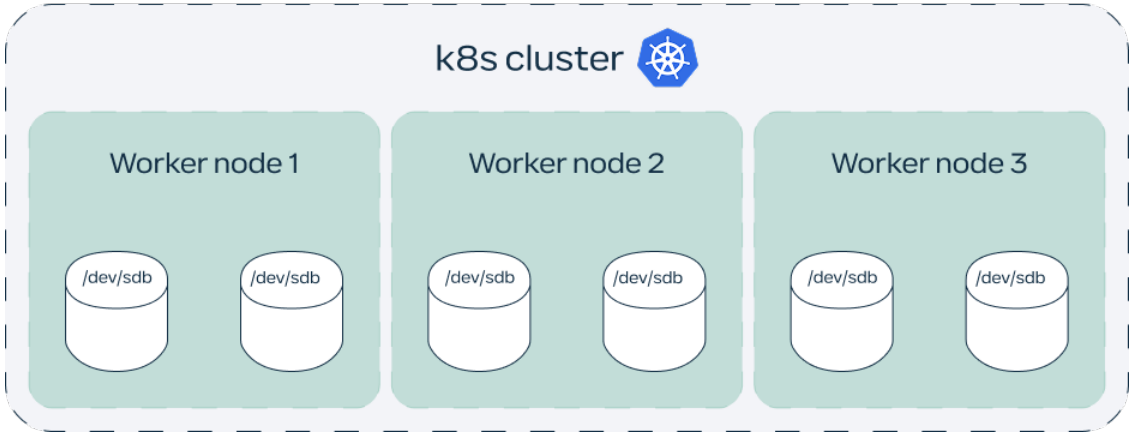
Аппаратный уровень

- NVIDIA GPU Operator (MIG / Time-slices)
- NVIDIA Network Operator (опционально)
- Longhorn

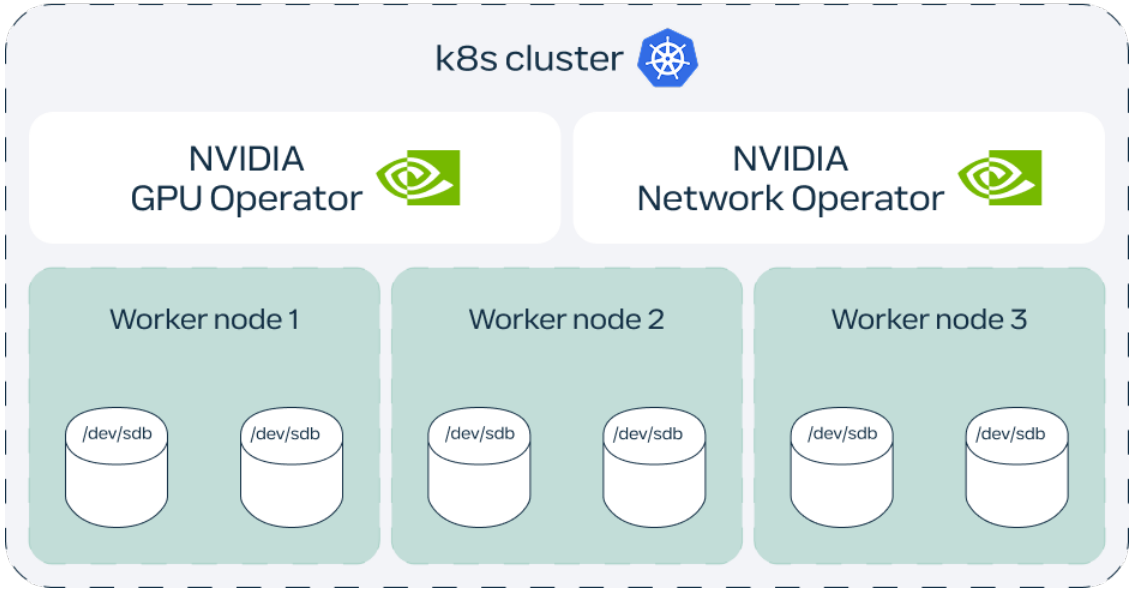
А что же есть в нашей песочнице?



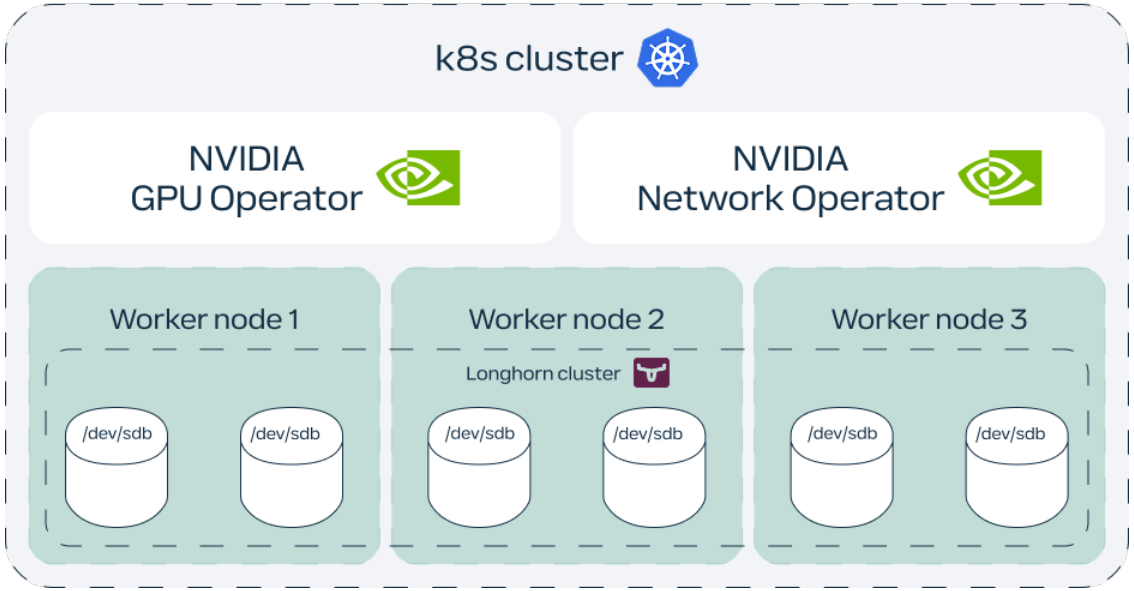
А что же есть в нашей песочнице?



А что же есть в нашей песочнице?



А что же есть в нашей песочнице?



А что же есть в нашей песочнице?

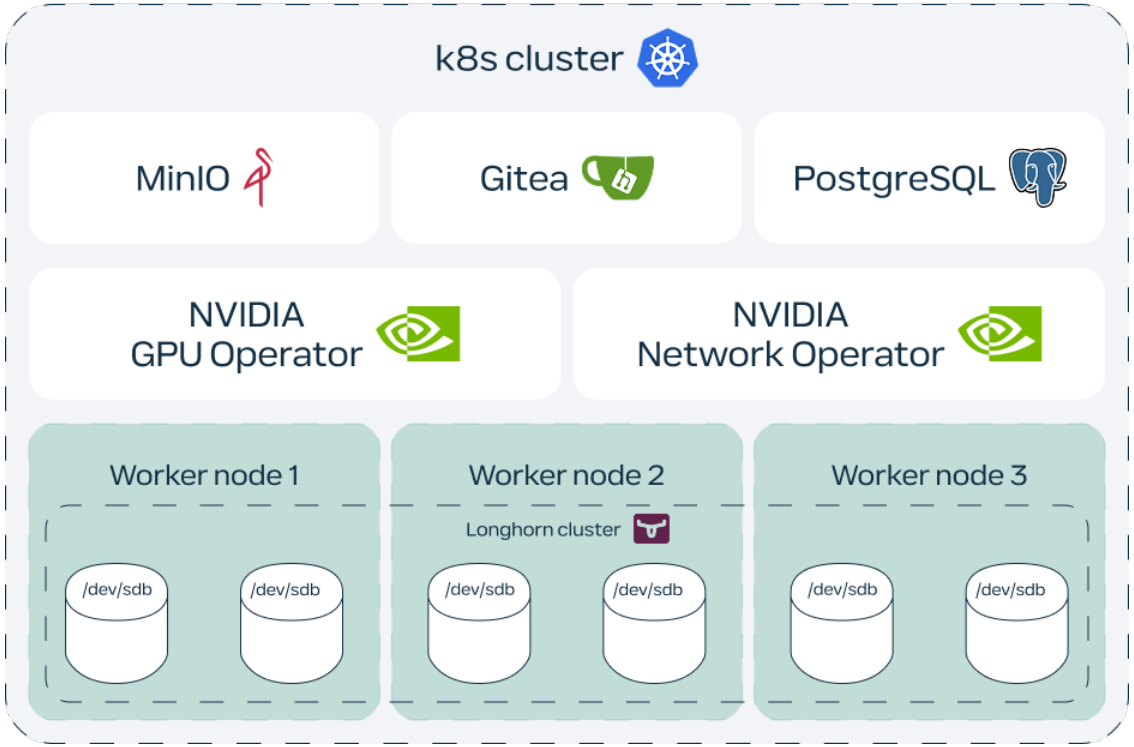
Аппаратный уровень

- NVIDIA GPU Operator (MIG / Time-slices)
- NVIDIA Network Operator (опционально)
- Longhorn

Хранилища

- MinIO
- PostgreSQL
- Gitea

А что же есть в нашей песочнице?



А что же есть в нашей песочнице?

Аппаратный уровень

- NVIDIA GPU Operator (MIG / Time-slices)
- NVIDIA Network Operator (опционально)
- Longhorn

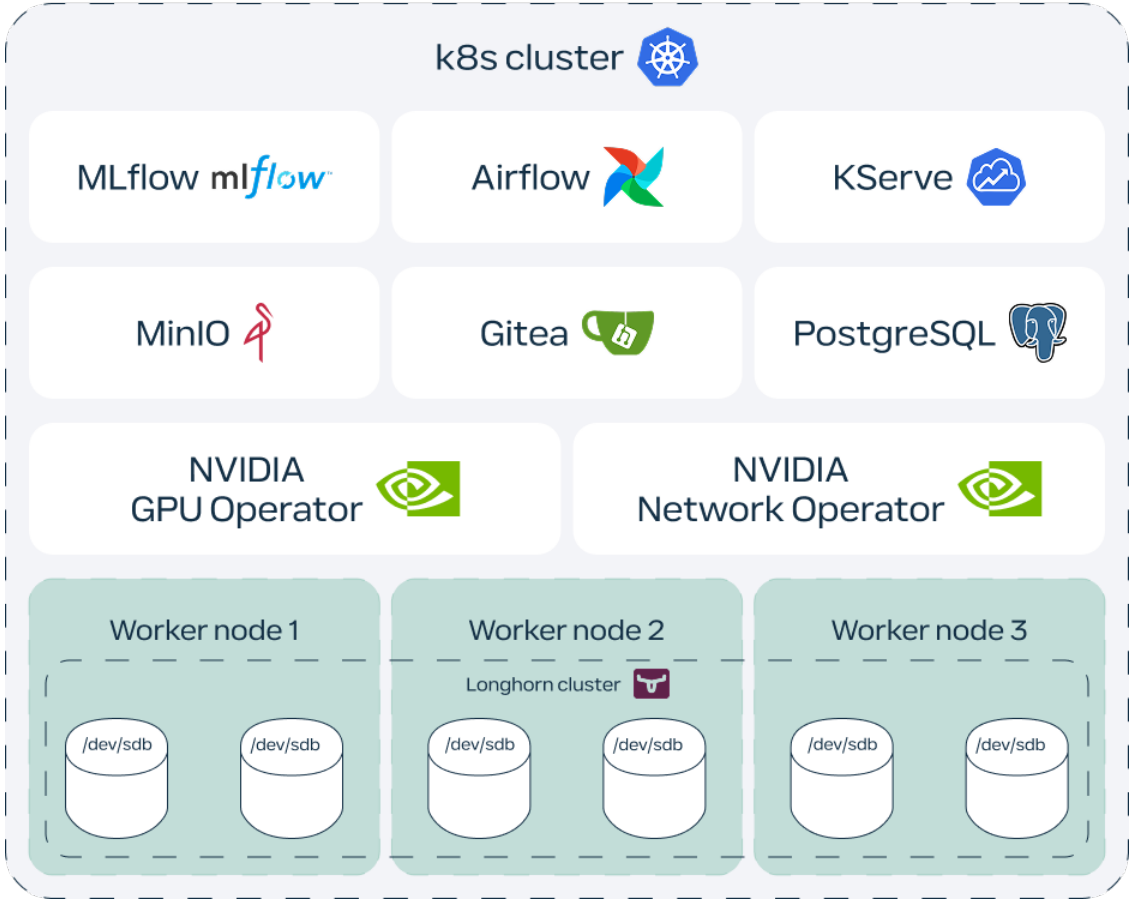
Хранилища

- ML-инженеры
- Data Scientists
- Архитекторы и инфраструктурные команды

Сервисный слой

- Mlflow
- Airflow
- KServe

А что же есть в нашей песочнице?



А что же есть в нашей песочнице?

Аппаратный уровень

- NVIDIA GPU Operator (MIG / Time-slices)
- NVIDIA Network Operator (опционально)
- Longhorn

Хранилища

- ML-инженеры
- Data Scientists
- Архитекторы и инфраструктурные команды

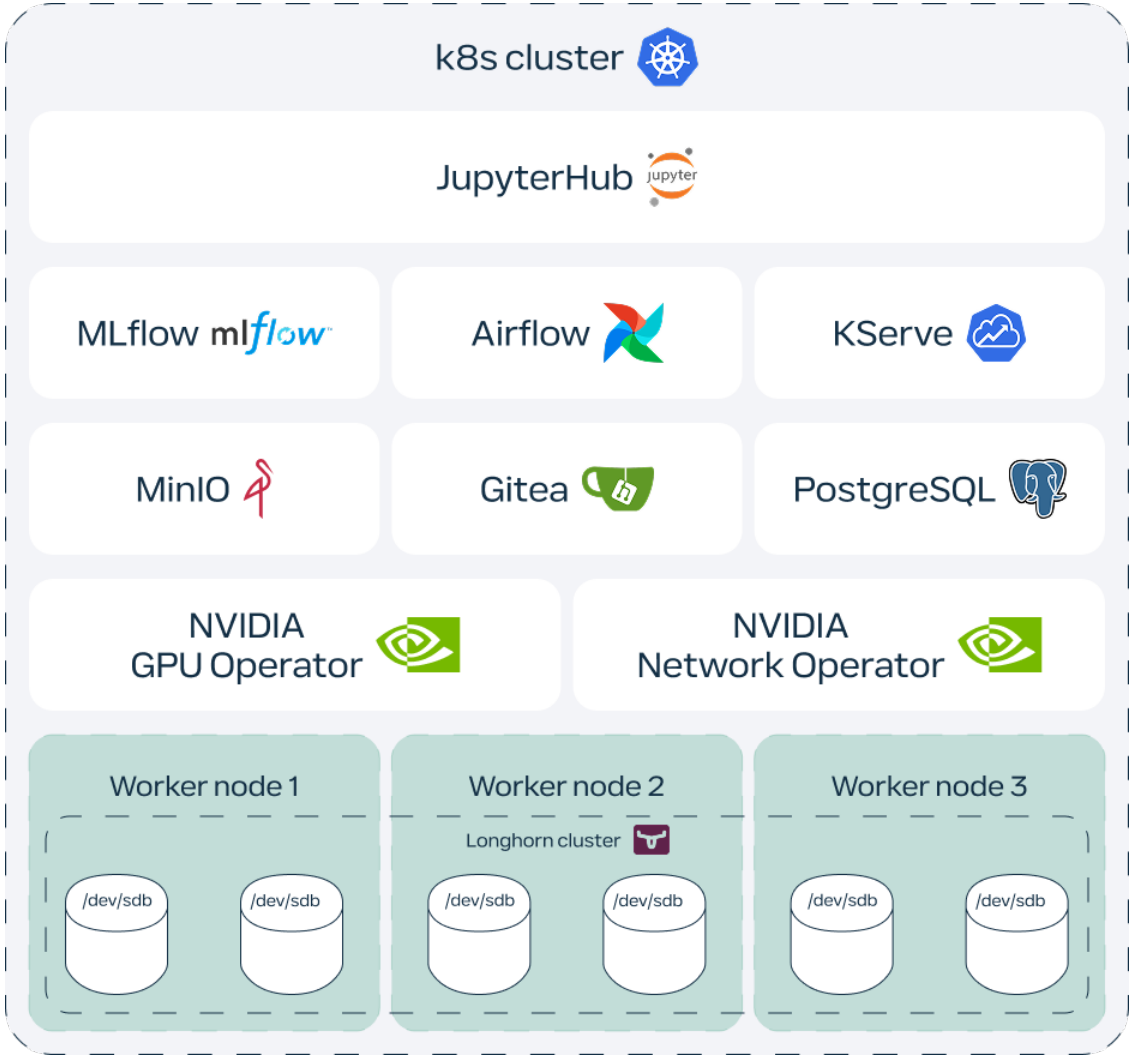
Сервисный слой

- Mlflow
- Airflow
- KServe

Интерфейс

- JupyterHub

А что же есть в нашей песочнице?



А что же есть в нашей песочнице?

Аппаратный уровень

- NVIDIA GPU Operator (MIG / Time-slices)
- NVIDIA Network Operator (опционально)
- Longhorn

Хранилища

- ML-инженеры
- Data Scientists
- Архитекторы и инфраструктурные команды

Сервисный слой

- Mlflow
- Airflow
- KServe

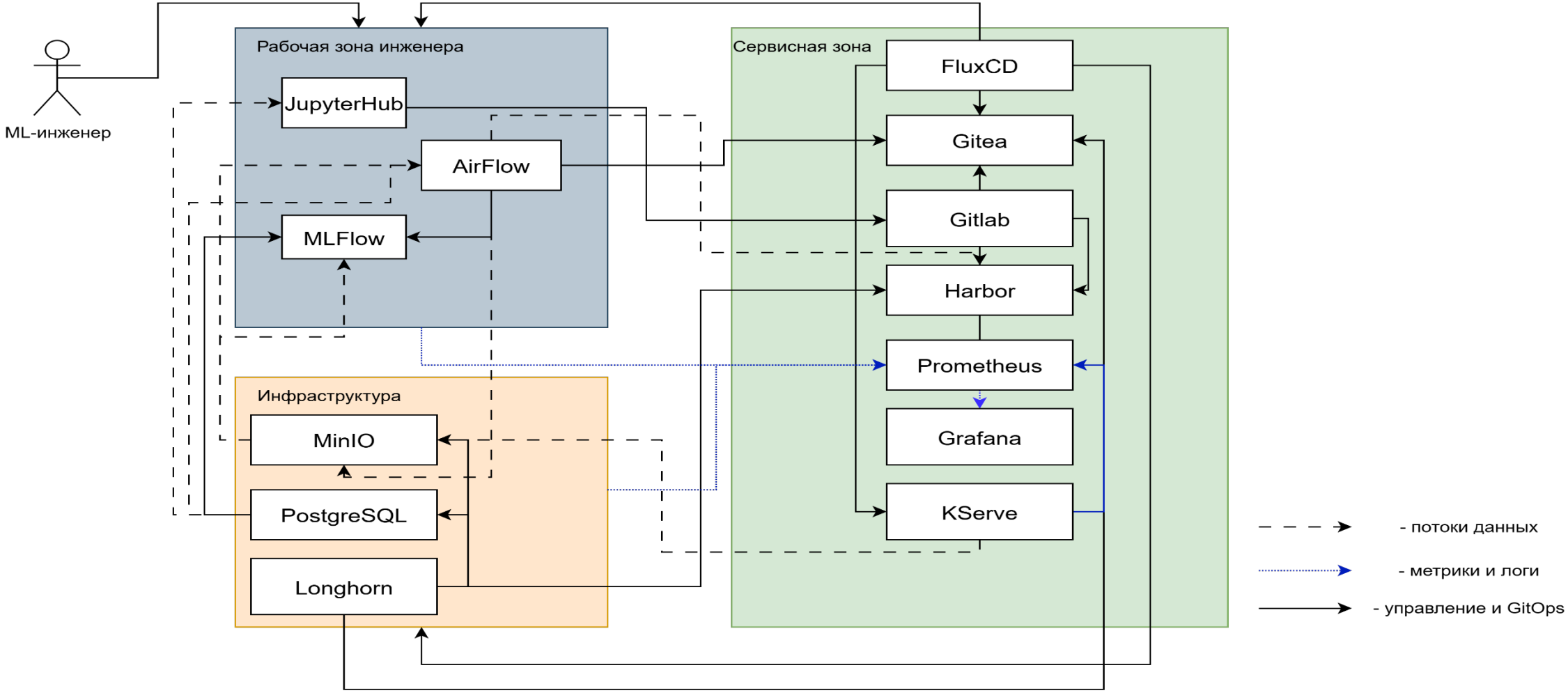
Интерфейс

- JupyterHub

Интеграции и автоматизация

- Harbor/Nexus
- GitLab

Архитектурная схема ML-платформы



Юху, практика!

Инфраструктура:

Платформа

k8s - Nova Container Platform

Воркеры

Два сервера Dell с NVIDIA T4

Внешний GitLab сервер

Задача

Классифицировать изображение

Датасет

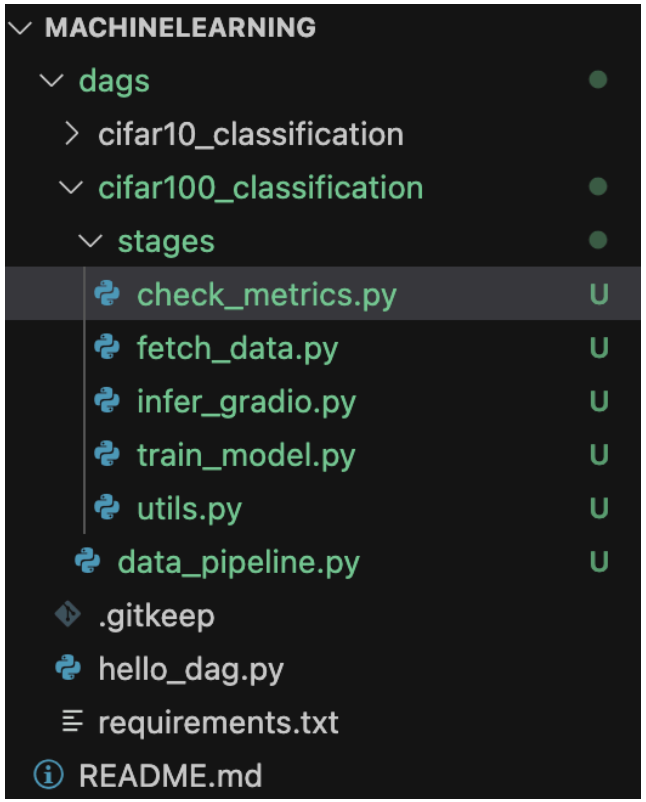
CIFAR100

Цель

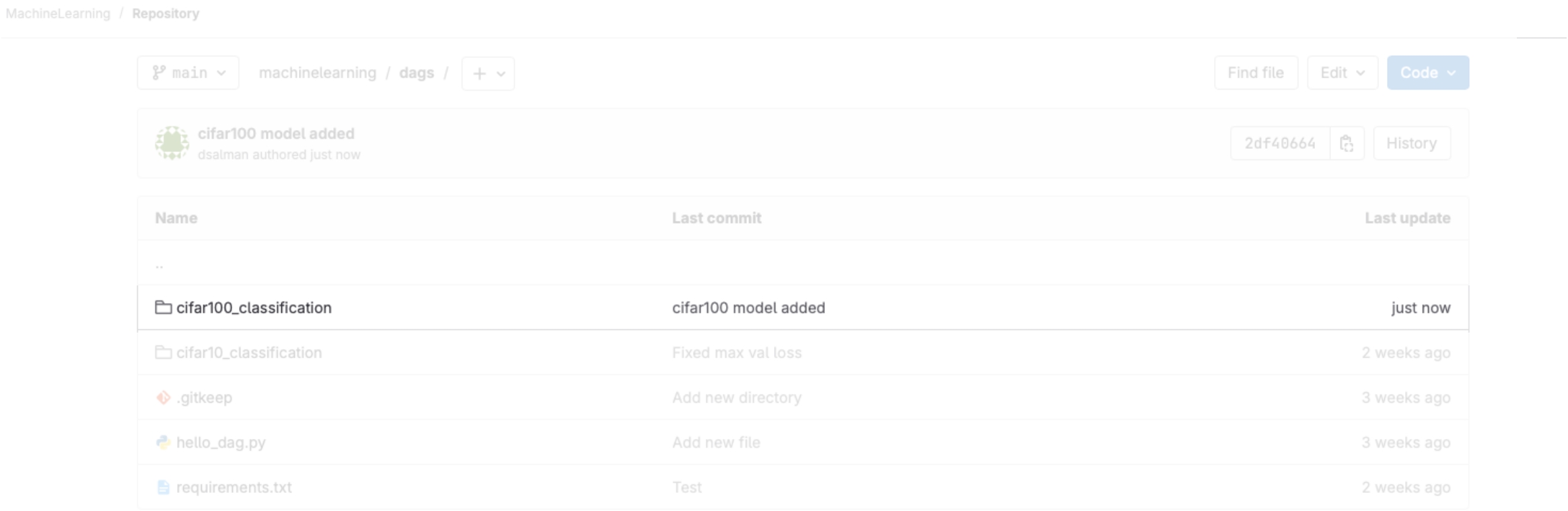
Бесшовно пройти весь процесс ML, не думая о бэкенде, где всё это работает (на самом деле я хочу узнать, как классифицирует моего кота ИИ)

Юху, практика!

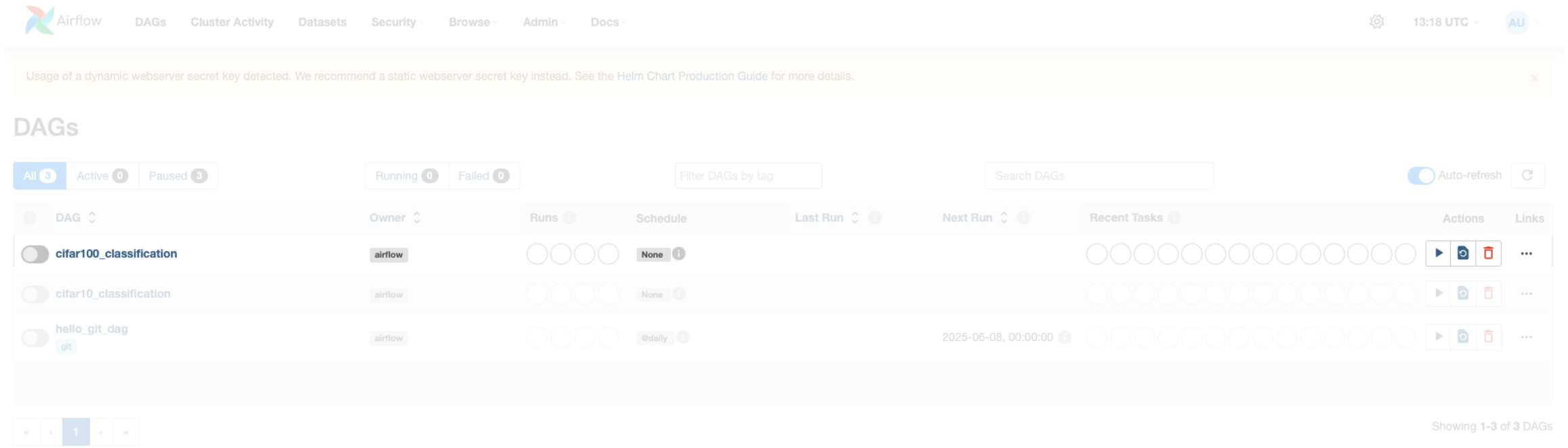
PUSH!



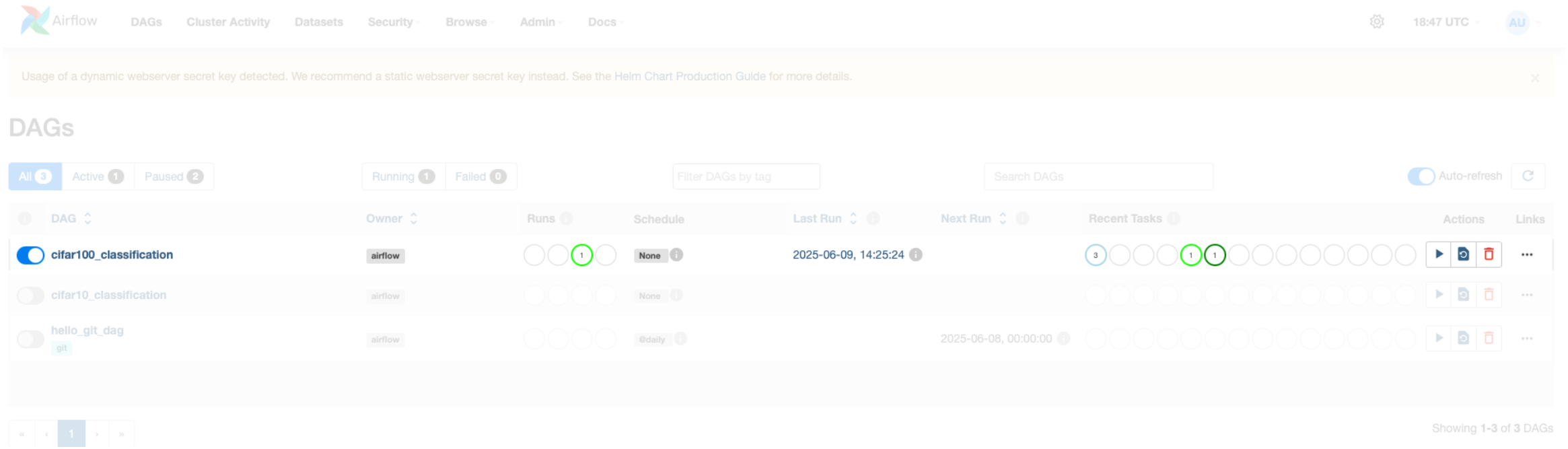
Юху, практика!



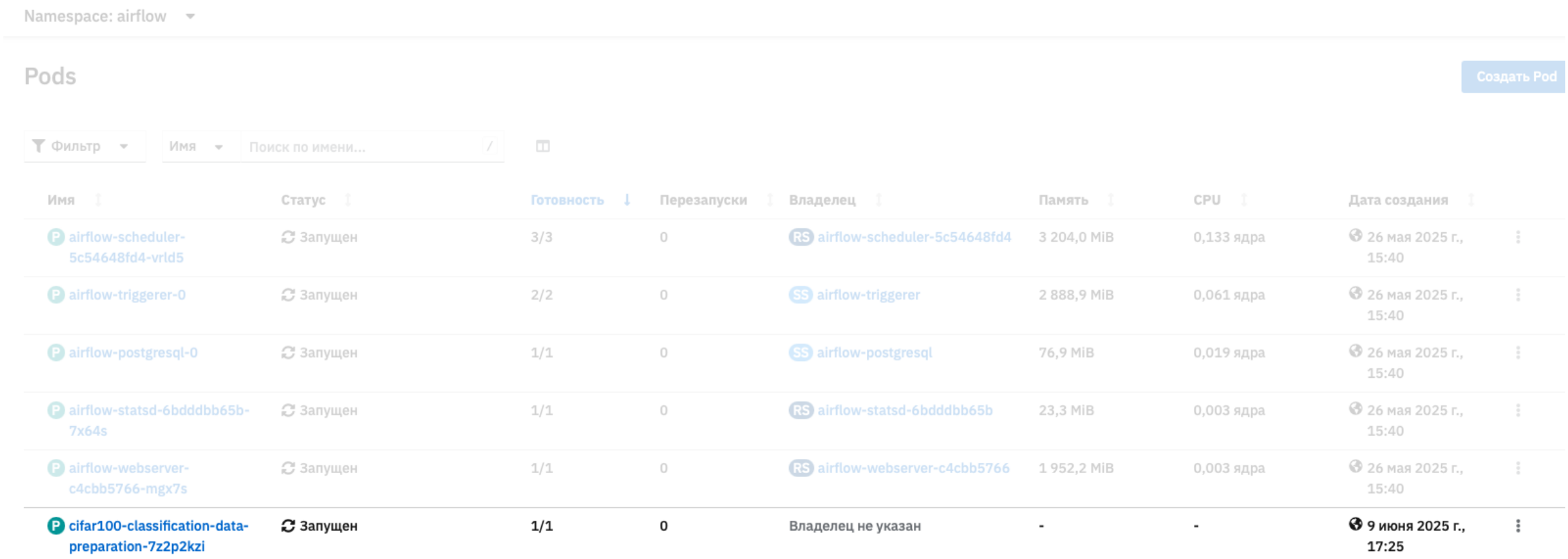
Юху, практика!



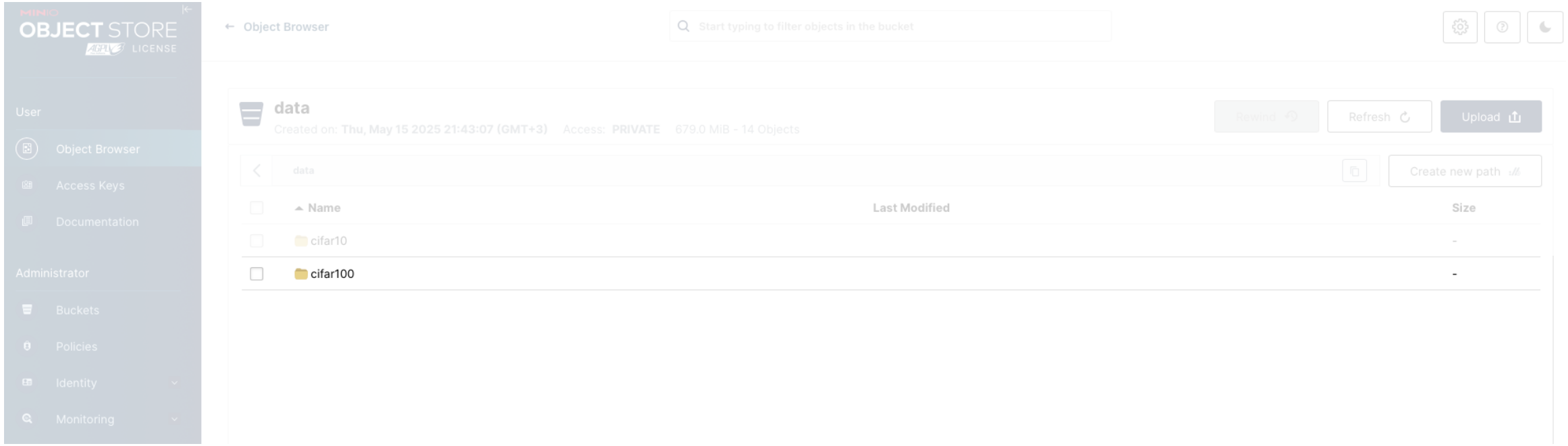
Юху, практика!



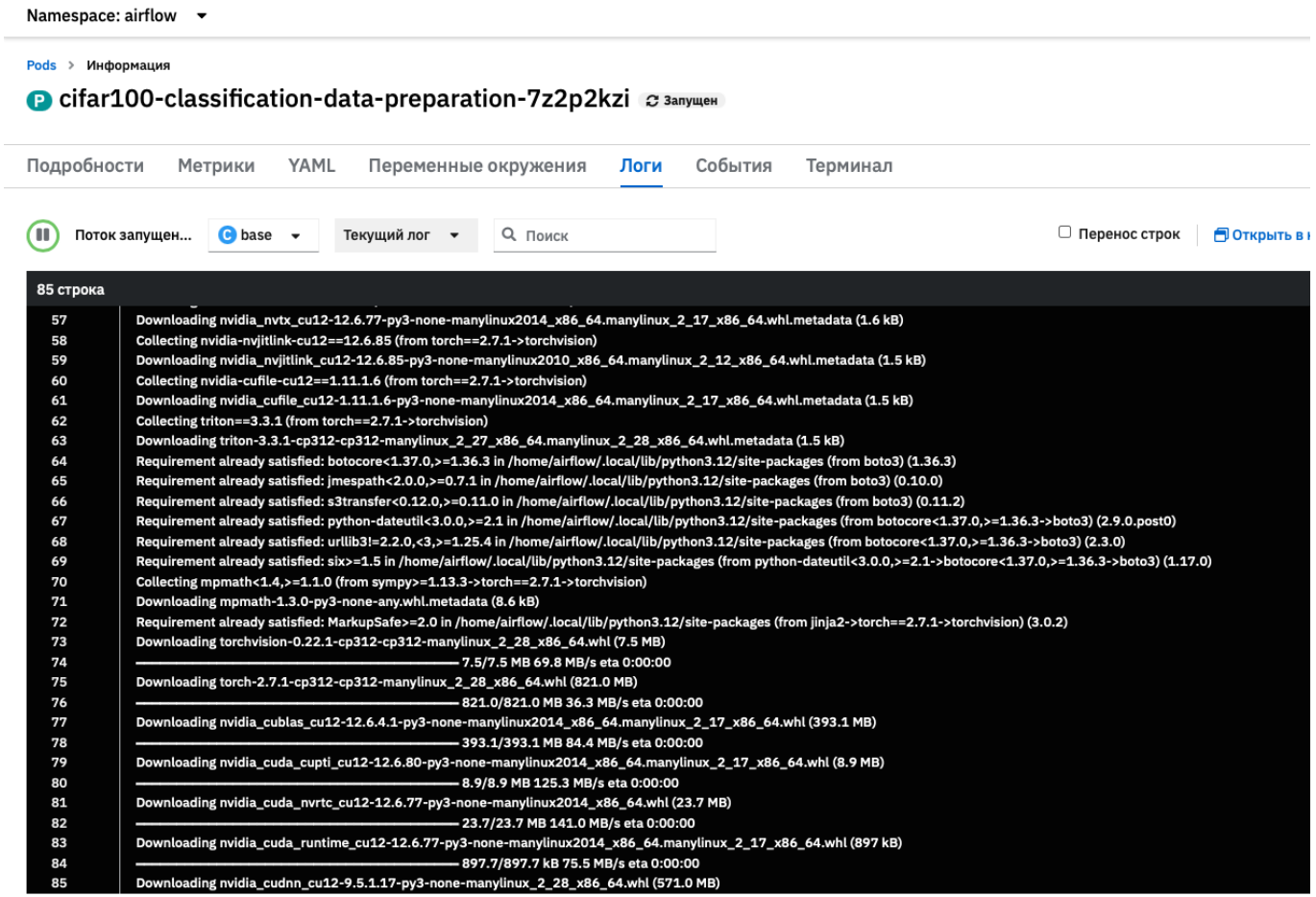
Юху, практика!



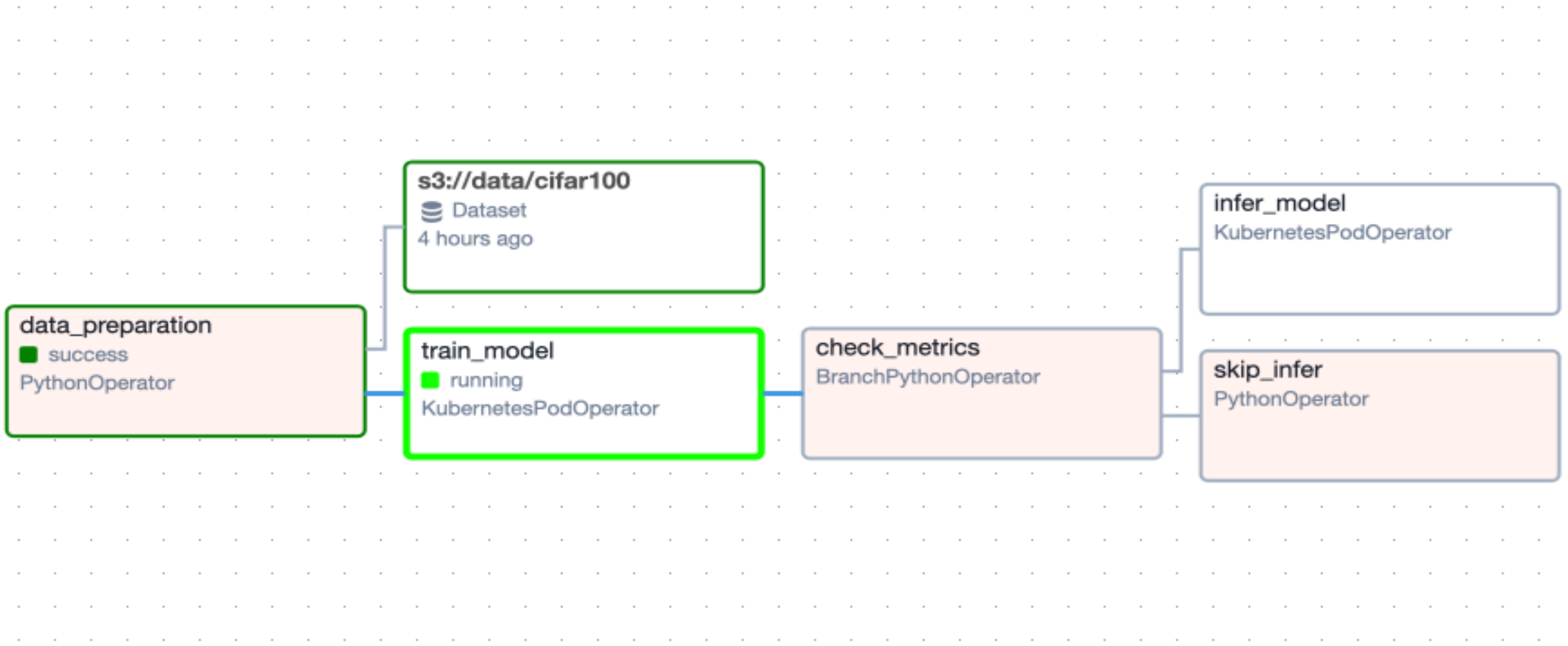
Юху, практика!



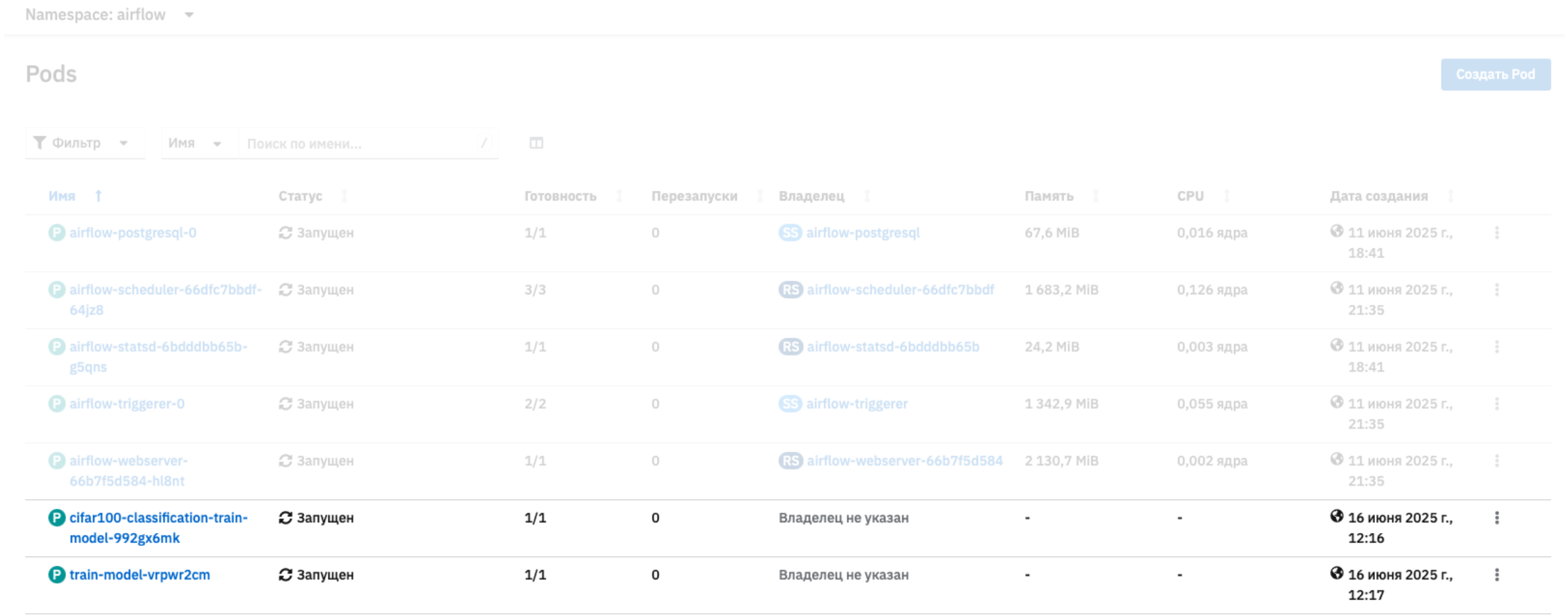
Юху, практика!



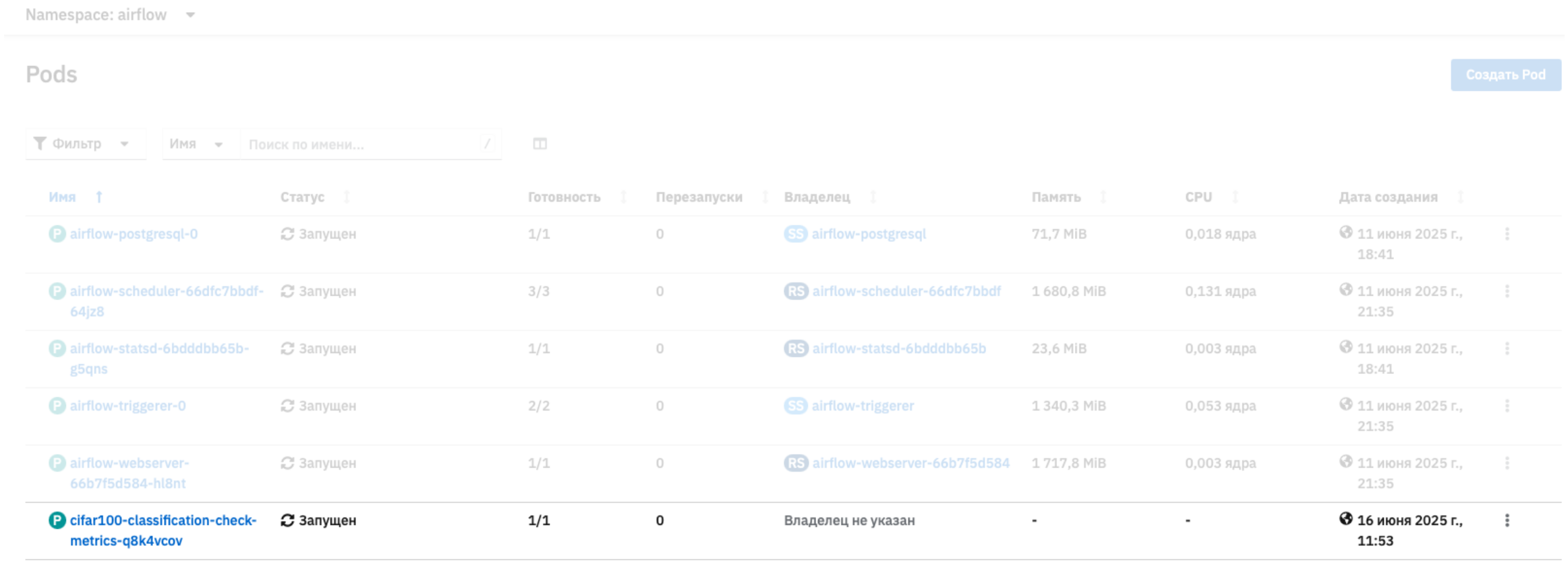
Юху, практика!



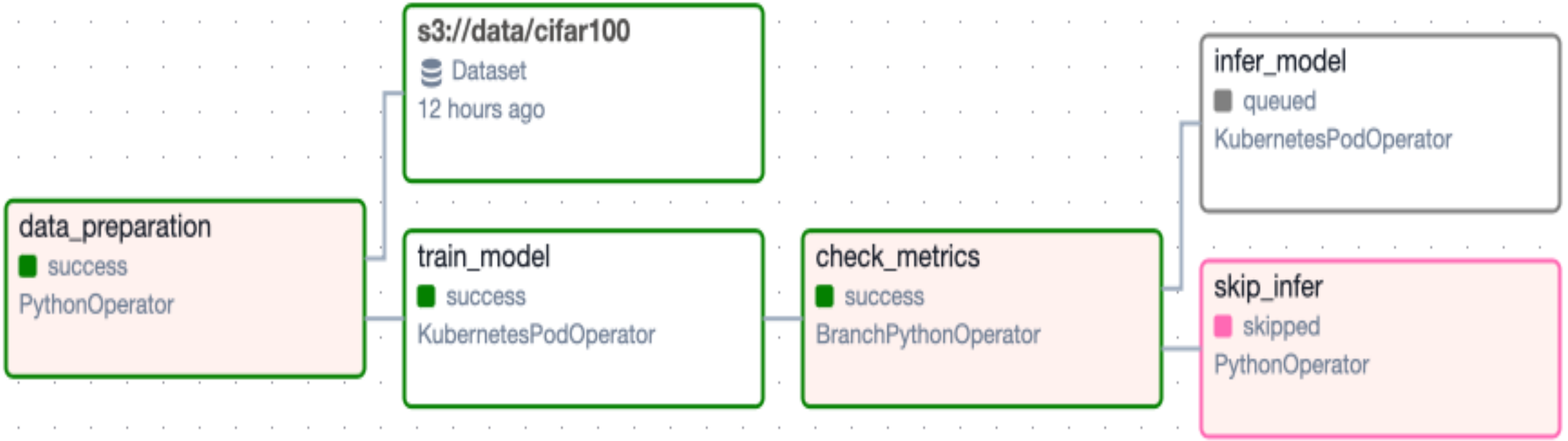
Юху, практика!



Юху, практика!



Юху, практика!



Юху, практика!

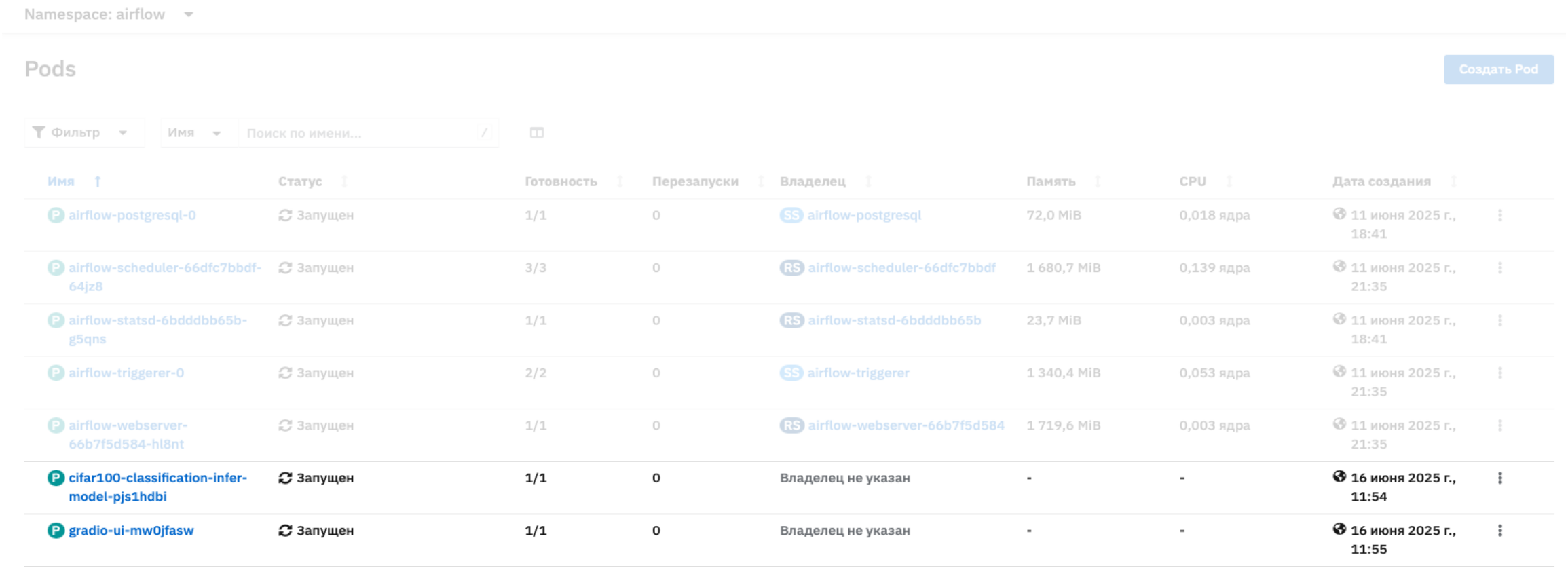
Манифест InferenceService
Дополнительно нужно создать две сущности:

- secret с s3-credentials
- service account с привязкой к этому секрету

```yaml
apiVersion: serving.kserve.io/v1beta1
kind: InferenceService
metadata:
  name: resnet18
  namespace: ml-inference
  annotations:
    serving.kserve.io/s3-endpoint: "minio.svc.cluster.local"
    serving.kserve.io/s3-credentials-secret-name: "s3-cred"
spec:
  predictor:
    serviceAccountName: resnet18-sa
    triton:
      runtimeVersion: "23.10-py3"
      storageUri: "s3://data/models"
      resources:
        requests:
          cpu: "500m"
          memory: "1Gi"
          nvidia.com/gpu: 1
        limits:
          cpu: "1"
          memory: "2Gi"
          nvidia.com/gpu: 1
    nodeSelector:
      nvidia.com/gpu.present: "true"
    tolerations:
      - key: "nvidia.com/gpu"
        operator: "Exists"
        effect: "NoSchedule"
```

https://github.com/kserve/kserve

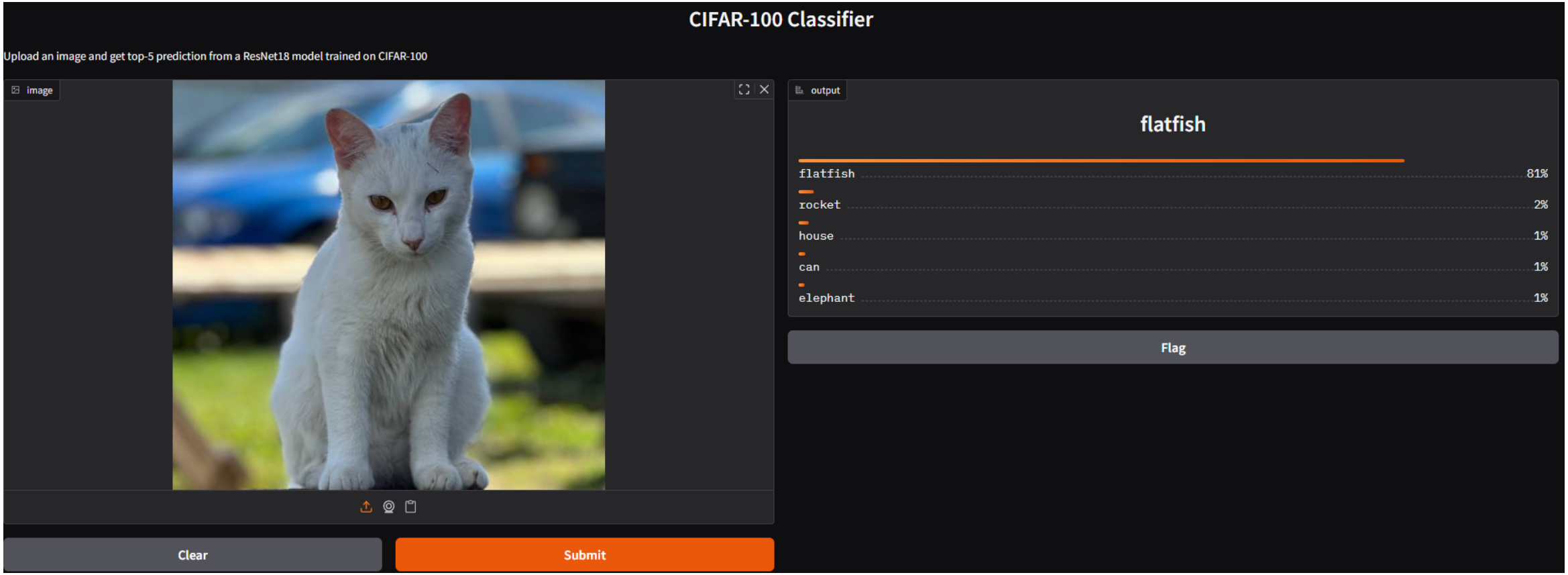
Юху, практика!



https://github.com/gradio-app/gradio

Юху, практика!

Юху, практика!



Юху, практика!

Metrics (6)

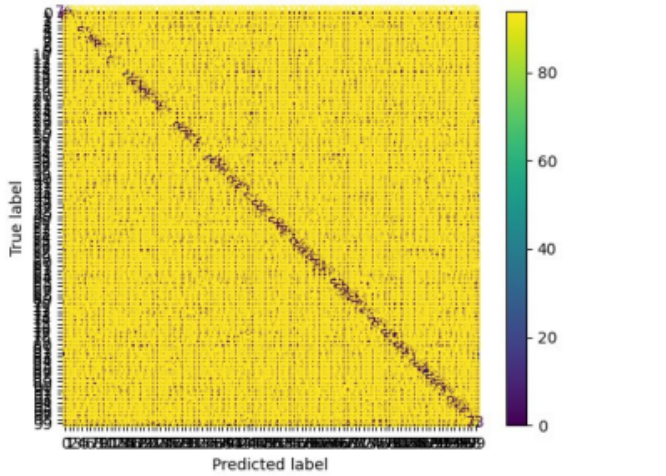
Search metrics

Metric	Value
val_loss	85.10773599147797
val_accuracy	0.5678
train_loss	96.06210631132126
val_precision	0.5675951905553892
val_recall	0.5672150327612843
val_f1_score	0.5646417947926783

Parameters (6)

Search parameters

Parameter	Value
dataset	cifar100
epochs	30
lr	0.001
model	ResNet18
batch_size	256
early_stopping_patience	20



✅ **Обучение модели завершено!**

🧠 **Model:** `ResNet18`
📅 **Timestamp:** `2025-06-16 19:52:16`
🔢 **Epochs:** `30` | **Batch:** `256` | **LR:** `0.001`
🖥️ **Device:** `cuda`

🎯 **Accuracy:** `0.5678`
📊 **F1-score:** `0.5646`
🔁 **Δ Accuracy:** `+0.0898`
🔁 **Δ F1-score:** `+0.0932`

🔗 Открыть в MLflow
`Run ID: cc80c2871f644d5fb051dc613faabd4b`

23:03

Почему песочница «идеальная»?

Time-to-first-demo ↓ в 5–7 раз
(с недель до дней)

Онбординг инженера ↓ с месяцев
до пары дней

GPU-utilization ↑ 30–40% за счёт MIG/
Time-slicing и автоскейлинга

Простой GPU ↓ ~50% → меньше
неиспользуемых карточек →
экономия бюджета

Time-to-production ↓ в 2–3 раза

99% экспериментов воспроизводимы
через MLflow

Время обнаружения инцидента ↓
с часов до минут

Reuse пайплайнов и экспериментов
↑ 60–70%

А какие еще «игрушки» можно затащить в песочницу?

Schedulers

KAI, Volcano

Инференс / деплой моделей

Triton, Dynamo, KServe

Визуализация и анализ данных

Grafana, Superset, Evidently AI

Операторы для работы с видеокартами других вендоров

HAMi, ROCm

Безопасность

SSO

Подпись образов (Cosign / Notary)

SBOM (Syft)

Сканирование уязвимостей (Trivy)

Управление секретами (Vault)

Правила запуска контейнеров (OPA Gatekeeper / Kyverno)

Настройка SSO для MLflow, Jupyterhub и Airflow

habr.com/ru/companies/orion_soft/articles/940592/

«Данила, план гениальный — как всегда!»
(с) Герман

source: Российское скетч-шоу «ДаЁшь молодЁжь!»

Выводы

Платформа является кросс-доменной, подходит для любой отрасли

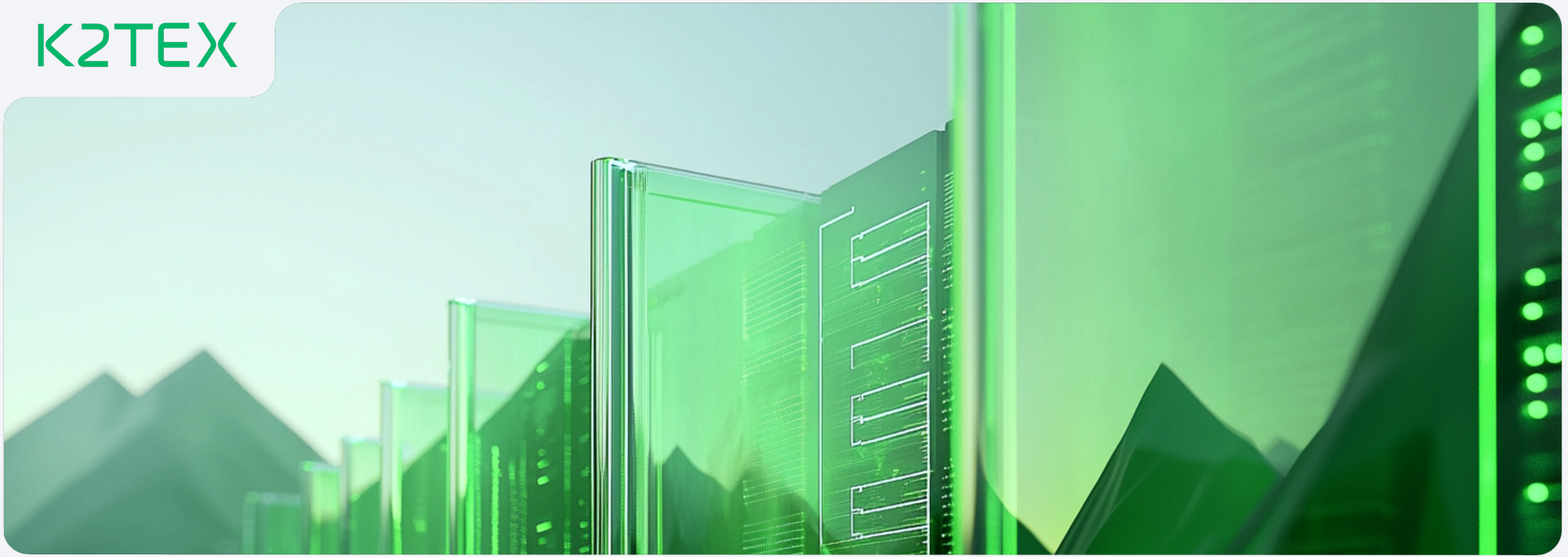
Эффективное использование рабочего времени и уменьшение ТРЗ на решение смежных проблем

Контроль ресурсов, в том числе GPU — больше не узкое место

Полноценный конвейер обкатки ML-модели больше не головная боль

Мониторинг встроен в процесс и помогает отслеживать качество и стабильность обкатывания моделей

K2TEX



Даниил Салман

Tech Lead по практике контейнеризации K2 Tech

tg: @DV_Salman

Будьте в курсе новостей K2 TEX

k2.tech
info@k2.tech
+7 (495) 797-85-84