

# Идеальная «песочница» для ML- моделей:

настраиваем контейнеризацию  
без стресса



**Даниил Салман**

Tech Lead по практике  
контейнеризации  
K2 Tech

K2TECH



Эксперты цифровизации бизнеса

**250+**

ведущих инженеров,  
senior-разработчиков  
и экспертов

**10+**

лет — средний стаж  
работы специалистов

**1 200+**

проектов в год  
в сфере ИТ-услуг

**9**

собственных  
ИТ-продуктов

**18**

лет на рынке

**1 500+**

ИТ-профессионалов

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...



Герман:



source: <https://is.gd/sDxhKK>

# Немного истории Германа...



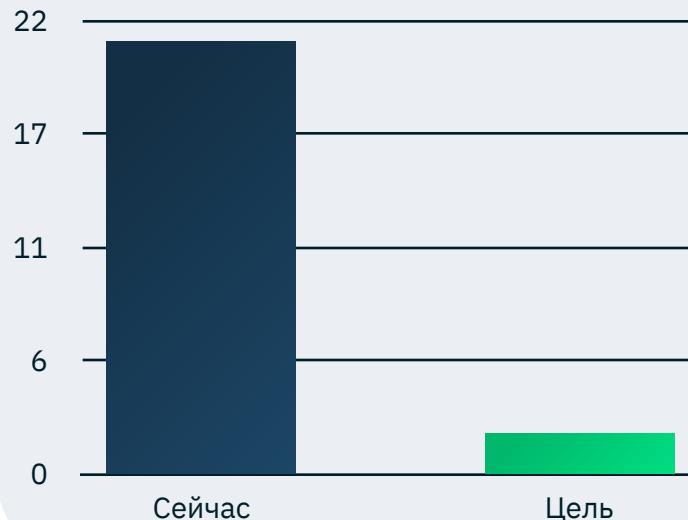


source: Российское скетч-шоу «ДаЁшь молодЁжь!»

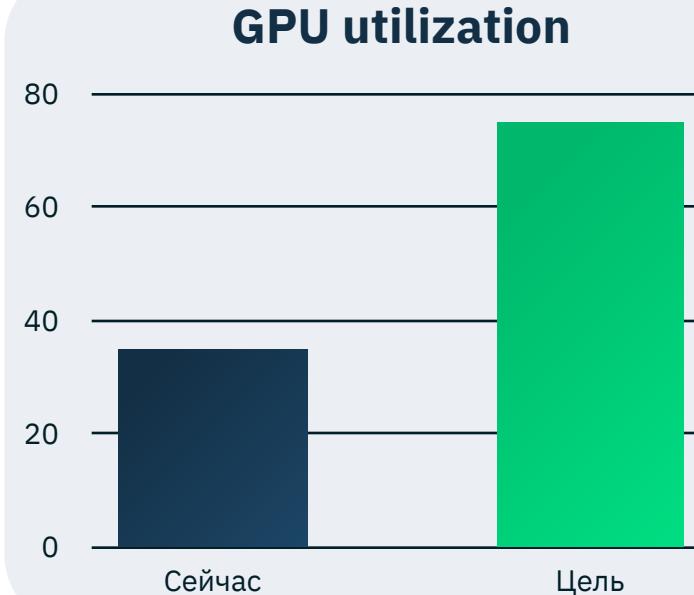
# С чем сталкиваются ML-инженеры

- ⌚ 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

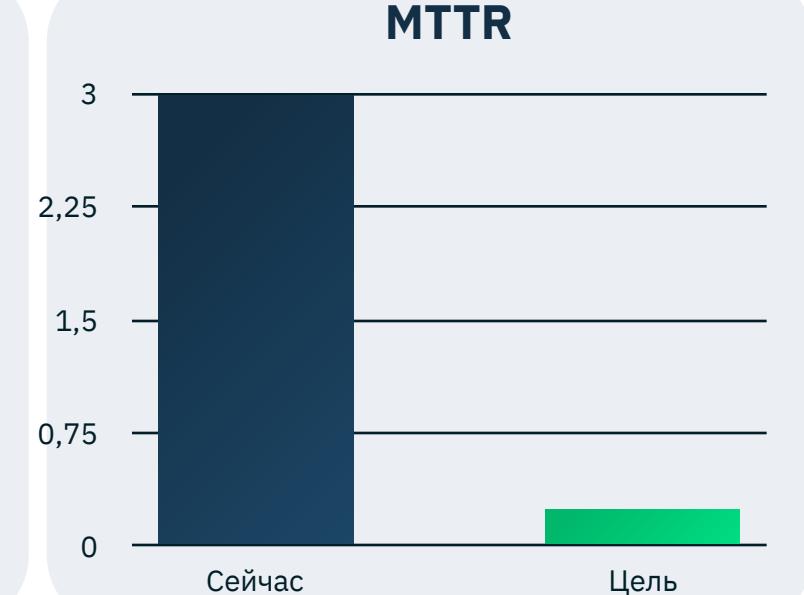
**Time-to-first-training**



**GPU utilization**



**MTTR**



# Подходы к развертыванию k8s

Способ	Для кого	Затраты	Контроль	Поддержка	Мониторинг	Плюсы	Минусы
<b>kubeadm</b>	Энтузиасты DevOps-гуру	Низкие	Максимальный	Сам	Свой	Максимальная гибкость, понимание изнутри	Много ручной работы
<b>kuberspray</b>	Рядовой DevOps-инженер	Средние	Высокий	Сообщество	Свой	Баланс гибкости и автоматизации, open source	Ansible
<b>KaaS в облаке</b>	Бизнес, готовый платить	Средние/высокие	Средний	Облако	Облако	Удобно, быстро, поддержка	Нет всех сервисов "из коробки", привязка к облаку
<b>Вендорские решения</b>	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	<ul style="list-style-type: none"><li>• Calico</li><li>• Cilium</li><li>• Flannel</li></ul>		<ul style="list-style-type: none"><li>• Containerd (default)</li><li>• Cri-o</li><li>• Docker</li></ul>
Service Discovery & Coordination	Service Proxy	Service Mesh	Backup	OS
CoreDNS etcd	kube-proxy			<ul style="list-style-type: none"><li>• Из коробки нет поддержки российских ОС</li></ul>
Conf. Management	Infrastructure providers	UI	K8S services	Scaling
			<p>kube-apiserver controller-manager scheduler etcd kubelet</p>	

# Technology stack NOVA

Observability	Security	Networking	Storage	Container Runtime
<ul style="list-style-type: none"> <li>Prometheus</li> <li>Grafana</li> <li>Alertmanager</li> <li>Thanos Query</li> <li>Logging Operator</li> <li>Opensearch</li> <li>Fluentd</li> </ul>	<ul style="list-style-type: none"> <li>RBAC</li> <li>Neuvector</li> <li>Secrets Webhook</li> <li>CertManager</li> <li>Secrets Manager (CSI, PKI, OAuth)</li> </ul>	Calico Cilium	<ul style="list-style-type: none"> <li>Longhorn</li> </ul>	Containerd
Service Discovery & Coordination	Service Proxy	Service Mesh	Backup	OS
CoreDNS etcd	kube-proxy		<ul style="list-style-type: none"> <li>Velero</li> </ul>	<ul style="list-style-type: none"> <li>RED OS</li> <li>AlmaLinux</li> <li>Astra Linux</li> <li>RHEL-based</li> </ul>
Conf. Management	Infrastructure providers	UI	K8S services	Scaling
<ul style="list-style-type: none"> <li>Gitea</li> <li>FluxCD</li> <li>Authomation tools</li> </ul>	<ul style="list-style-type: none"> <li>Bare Metal</li> <li>oVirt</li> <li>zVirt</li> <li>VMware vSphere</li> </ul>	<ul style="list-style-type: none"> <li>Nova Console</li> </ul>	<ul style="list-style-type: none"> <li>kube-apiserver</li> <li>controller-manager</li> <li>scheduler</li> <li>etcd</li> <li>kubelet</li> </ul>	<ul style="list-style-type: none"> <li>VPA</li> <li>HPA</li> </ul>

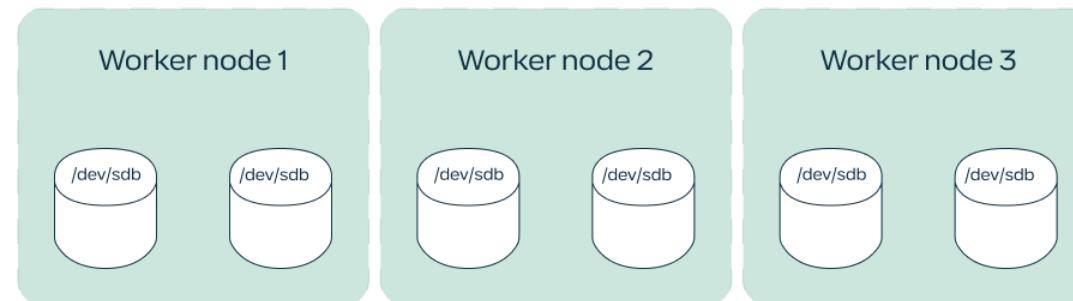


# А что же есть в нашей песочнице?

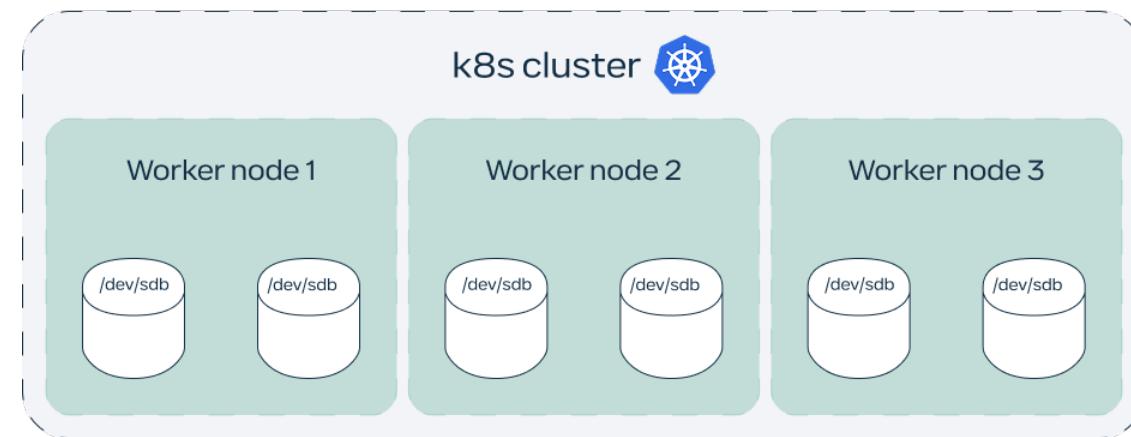
## Аппаратный уровень

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

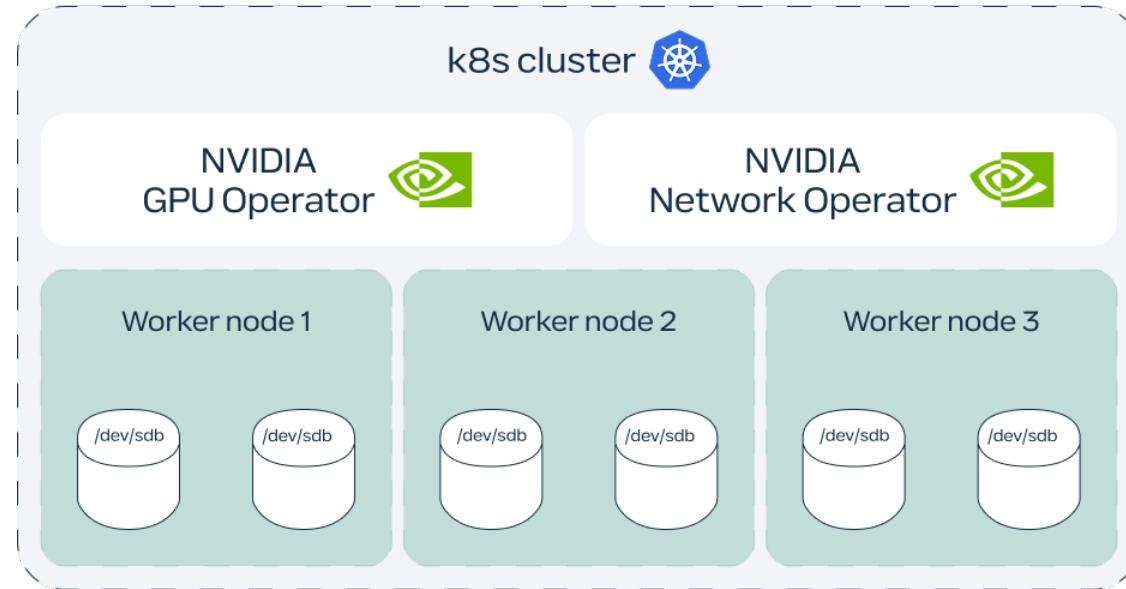
# А что же есть в нашей песочнице?



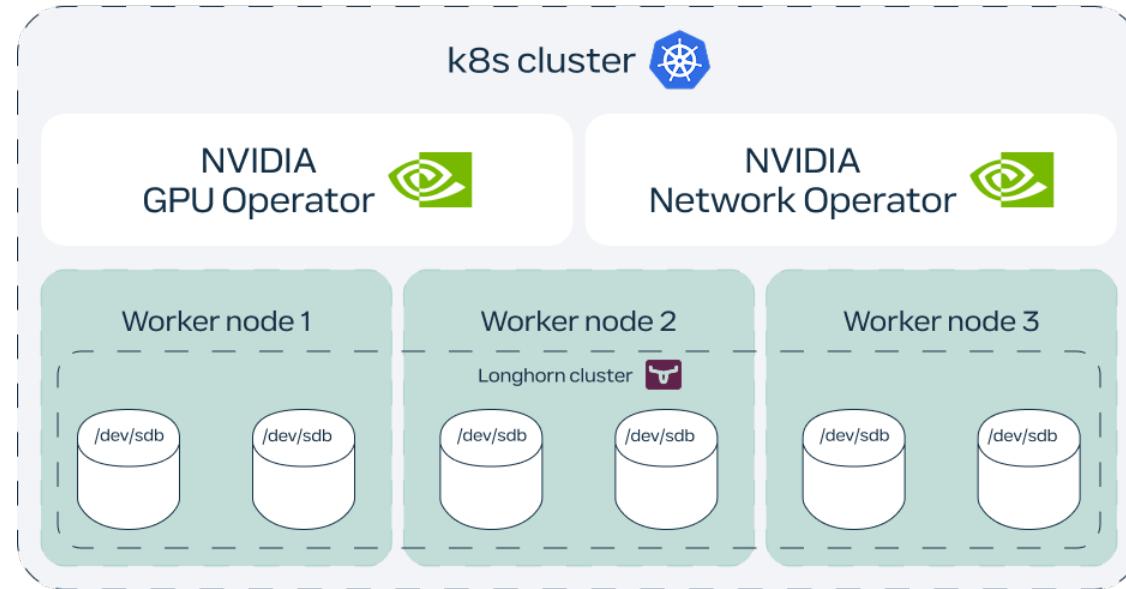
# А что же есть в нашей песочнице?



# А что же есть в нашей песочнице?



# А что же есть в нашей песочнице?



# А что же есть в нашей песочнице?

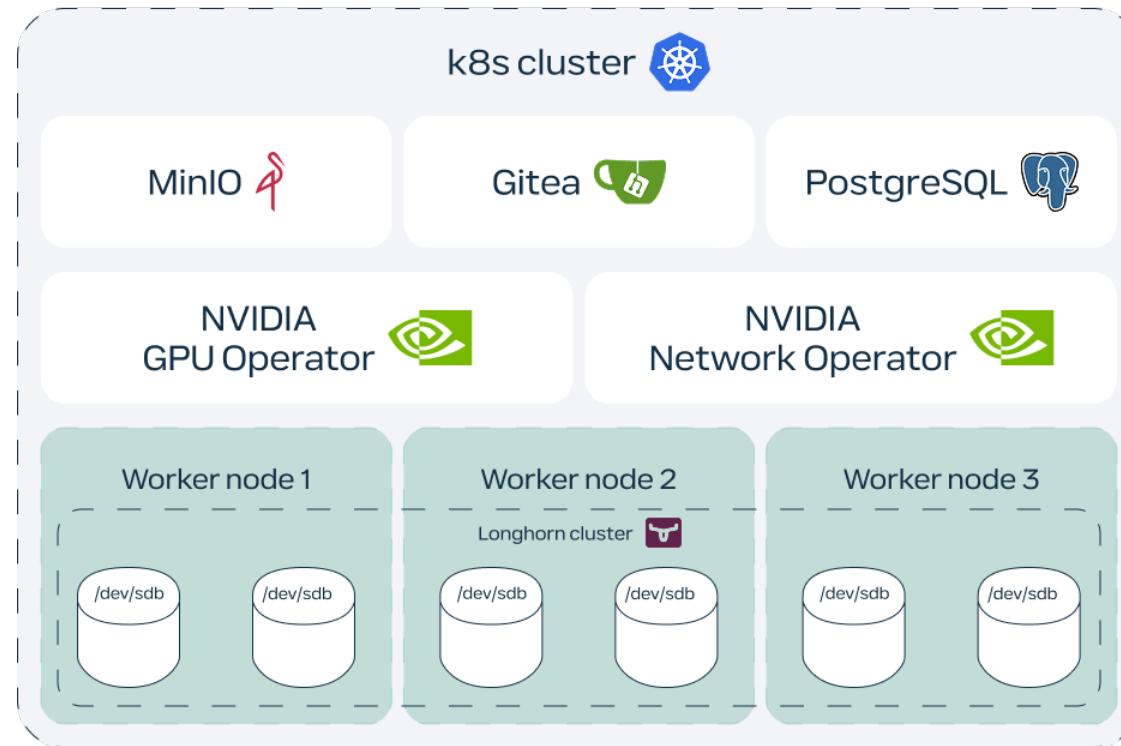
## Аппаратный уровень

- 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

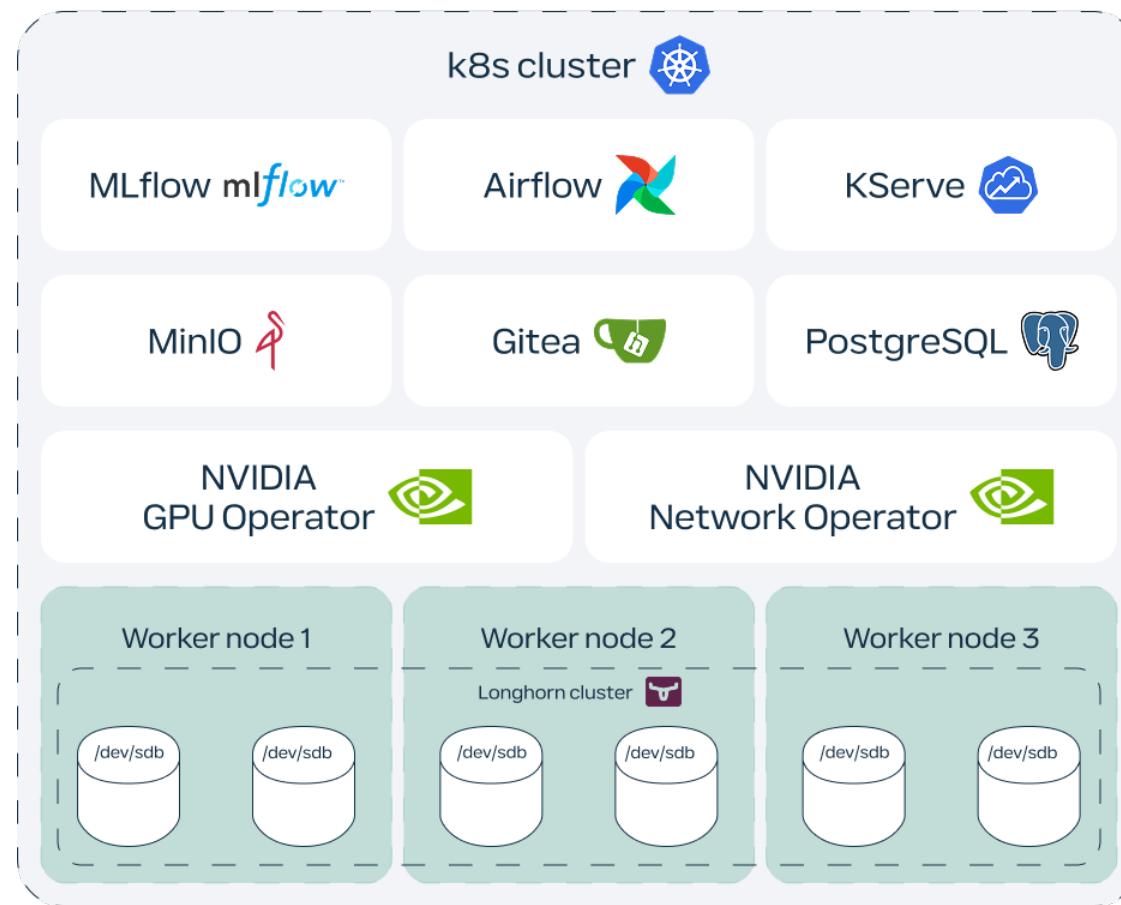
## Хранилища

- 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

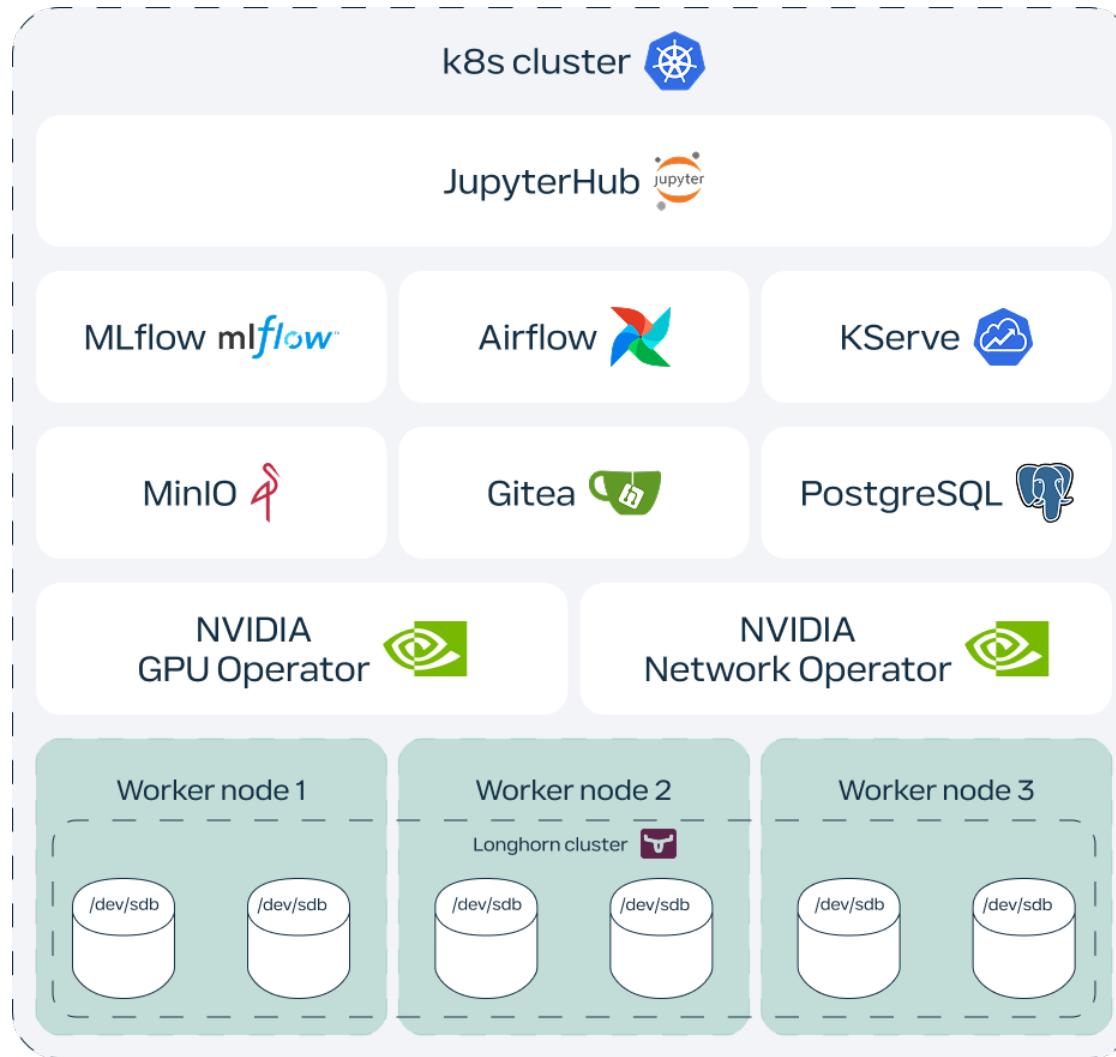
## Сервисный слой

- 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

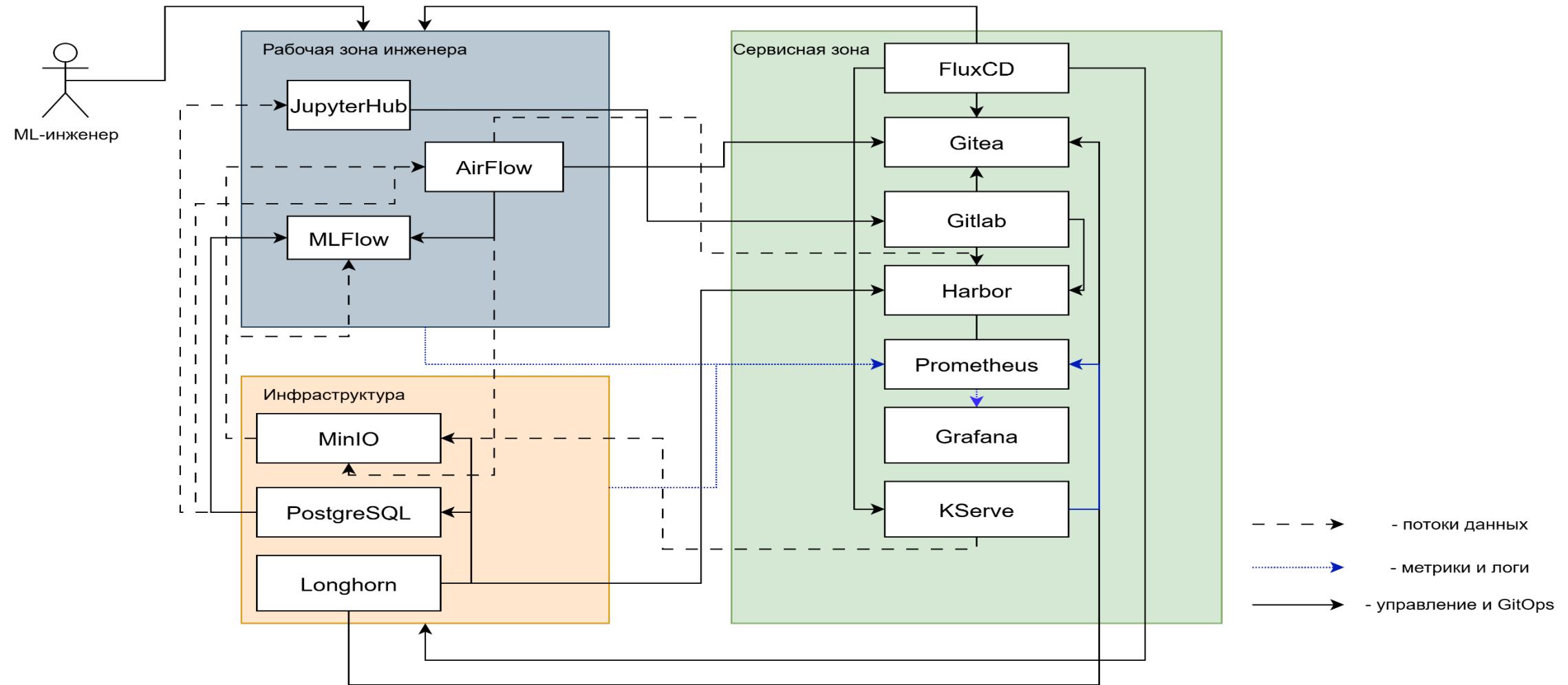
## Интерфейс

- 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!

```
✓ MACHINELEARNING
  ✓ dags
    > cifar10_classification
    ✓ cifar100_classification
      ✓ stages
        • check_metrics.py          U
        • fetch_data.py            U
        • infer_gradio.py          U
        • train_model.py           U
        • utils.py                 U
        • data_pipeline.py         U
      ◊ .gitkeep
      • hello_dag.py
      ≡ requirements.txt
  ⓘ README.md
```

# Юху, практика!

MachineLearning / Repository

Name	Last commit	Last update
..		
📁 cifar100_classification	cifar100 model added	just now
📁 cifar10_classification	Fixed max val loss	2 weeks ago
↳ .gitkeep	Add new directory	3 weeks ago
🐍 hello_dag.py	Add new file	3 weeks ago
📄 requirements.txt	Test	2 weeks ago

# Юху, практика!

The screenshot shows the Airflow web interface with the following details:

- Header:** Airflow logo, navigation links (DAGs, Cluster Activity, Datasets, Security, Browse, Admin, Docs), timestamp (13:18 UTC), and user info (AU).
- Message:** "Usage of a dynamic webserver secret key detected. We recommend a static webserver secret key instead. See the Helm Chart Production Guide for more details."
- Section Header:** DAGs
- Filter and Search:** Buttons for All (3), Active (0), Paused (3), Running (0), Failed (0). Filter DAGs by tag and Search DAGs input fields. Auto-refresh and refresh buttons.
- DAG List:** A table listing three DAGs:
  - cifar100\_classification**: Owner airflow, Runs 5, Schedule None, Last Run N/A, Next Run N/A, Recent Tasks 15, Actions: play, download, delete, more.
  - cifar10\_classification**: Owner airflow, Runs 3, Schedule None, Last Run N/A, Next Run N/A, Recent Tasks 15, Actions: play, download, delete, more.
  - hello\_git\_dag**: Owner git, Runs 5, Schedule @daily, Last Run 2025-06-08, 00:00:00, Recent Tasks 15, Actions: play, download, delete, more.
- Pagination:** Page number 1 of 1.
- Text:** Showing 1-3 of 3 DAGs.

# Юху, практика!

The screenshot shows the Airflow web interface with the following details:

- Header:** Airflow logo, navigation links (DAGs, Cluster Activity, Datasets, Security, Browse, Admin, Docs), and a timestamp (18:47 UTC) with a dropdown.
- Message Bar:** "Usage of a dynamic webserver secret key detected. We recommend a static webserver secret key instead. See the Helm Chart Production Guide for more details." with a close button.
- Title:** DAGs
- Filtering and Search:** Buttons for All (3), Active (1), Paused (2), Running (1), Failed (0), Filter DAGs by tag, and Search DAGs. An Auto-refresh toggle and a refresh icon are also present.
- DAG List:** A table listing three DAGs:
  - cifar100\_classification**: Owner airflow, 1 run (green circle), None schedule, Last Run 2025-06-09, 14:25:24, 3 recent tasks (green circles). Status: Active.
  - cifar10\_classification**: Owner airflow, 0 runs, None schedule, Last Run N/A, 0 recent tasks. Status: Paused.
  - hello\_git\_dag**: Owner airflow, git, 0 runs, @daily schedule, Last Run 2025-06-08, 00:00:00, 0 recent tasks. Status: Paused.
- Pagination:** Page number 1 of 1.
- Page Footer:** Showing 1-3 of 3 DAGs.

# Юху, практика!

Namespace: airflow ▾

### Pods

Создать Pod

Фильтр	Имя	Статус	Готовность	Перезапуски	Владелец	Память	CPU	Дата создания	⋮
	airflow-scheduler-5c54648fd4-vrld5	⌚ Запущен	3/3	0	RS airflow-scheduler-5c54648fd4	3 204,0 MiB	0,133 ядра	⌚ 26 мая 2025 г., 15:40	⋮
	airflow-triggerer-0	⌚ Запущен	2/2	0	SS airflow-triggerer	2 888,9 MiB	0,061 ядра	⌚ 26 мая 2025 г., 15:40	⋮
	airflow-postgresql-0	⌚ Запущен	1/1	0	SS airflow-postgresql	76,9 MiB	0,019 ядра	⌚ 26 мая 2025 г., 15:40	⋮
	airflow-statsd-6bdddbb65b-7x64s	⌚ Запущен	1/1	0	RS airflow-statsd-6bdddbb65b	23,3 MiB	0,003 ядра	⌚ 26 мая 2025 г., 15:40	⋮
	airflow-webserver-c4ccb5766-mgx7s	⌚ Запущен	1/1	0	RS airflow-webserver-c4ccb5766	1 952,2 MiB	0,003 ядра	⌚ 26 мая 2025 г., 15:40	⋮
	cifar100-classification-data-preparation-7z2p2kzi	⌚ Запущен	1/1	0	Владелец не указан	-	-	⌚ 9 июня 2025 г., 17:25	⋮

# Юху, практика!

MINIO  
**OBJECT STORE**  
ASPLv2 LICENSE

User

- Object Browser
- Access Keys
- Documentation

Administrator

- Buckets
- Policies
- Identity
- Monitoring

← Object Browser

Start typing to filter objects in the bucket

data

Created on: Thu, May 15 2025 21:43:07 (GMT+3) Access: PRIVATE 679.0 MiB - 14 Objects

Rewind ⏪ Refresh ⏴ Upload ⬆

Name Last Modified Size

Name	Last Modified	Size
cifar10		-
cifar100		-

Create new path ⌂

# Юху, практика!

Namespace: airflow ▾

Pods > Информация

⟳ cifar100-classification-data-preparation-7z2p2kzi ⚡ Запущен

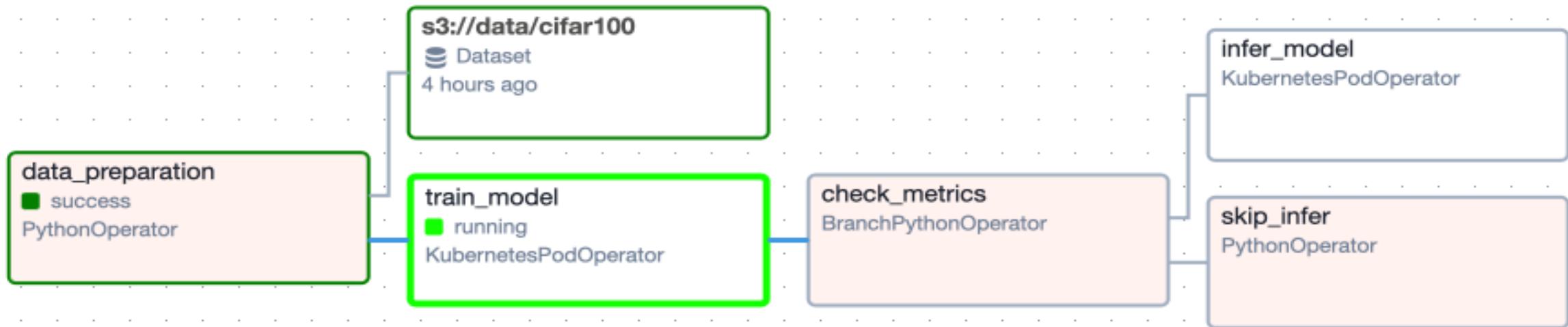
Подробности Метрики YAML Переменные окружения Логи События Терминал

⏸ Поток запущен... C base Текущий лог Поиск Перенос строк Открыть в i

85 строка

```
57 | Downloading nvidia_nvtx_cu12-12.6.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.6 kB)
58 | Collecting nvidia_nvjitlink_cu12==12.6.85 (from torch==2.7.1->torchvision)
59 | Downloading nvidia_nvjitlink_cu12-12.6.85-py3-none-manylinux2010_x86_64.manylinux_2_12_x86_64.whl.metadata (1.5 kB)
60 | Collecting nvidia_cufile_cu12==1.11.1.6 (from torch==2.7.1->torchvision)
61 | Downloading nvidia_cufile_cu12-1.11.1.6-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl.metadata (1.5 kB)
62 | Collecting triton==3.3.1 (from torch==2.7.1->torchvision)
63 | Downloading triton-3.3.1-cp312-manylinux_2_27_x86_64.manylinux_2_28_x86_64.whl.metadata (1.5 kB)
64 | Requirement already satisfied: botocore<1.37.0,>=1.36.3 in /home/airflow/.local/lib/python3.12/site-packages (from boto3) (1.36.3)
65 | Requirement already satisfied: jmespath<2.0.0,>=0.7.1 in /home/airflow/.local/lib/python3.12/site-packages (from boto3) (0.10.0)
66 | Requirement already satisfied: s3transfer<0.12.0,>=0.11.0 in /home/airflow/.local/lib/python3.12/site-packages (from boto3) (0.11.2)
67 | Requirement already satisfied: python-dateutil<3.0.0,>=2.1 in /home/airflow/.local/lib/python3.12/site-packages (from botocore<1.37.0,>=1.36.3->boto3) (2.9.0.post0)
68 | Requirement already satisfied: urllib3<2.2.0,<3,>=1.25.4 in /home/airflow/.local/lib/python3.12/site-packages (from botocore<1.37.0,>=1.36.3->boto3) (2.3.0)
69 | Requirement already satisfied: six=<1.5 in /home/airflow/.local/lib/python3.12/site-packages (from python-dateutil<3.0.0,>=2.1->botocore<1.37.0,>=1.36.3->boto3) (1.17.0)
70 | Collecting mpmath<1.4,>=1.1.0 (from sympy==1.13.3->torch==2.7.1->torchvision)
71 | Downloading mpmath-1.3.0-py3-none-any.whl.metadata (8.6 kB)
72 | Requirement already satisfied: MarkupSafe==2.0 in /home/airflow/.local/lib/python3.12/site-packages (from jinja2>=2.7.1->torchvision) (3.0.2)
73 | Downloading torchvision-0.22.1-cp312-cp312-manylinux_2_28_x86_64.whl (7.5 MB)
74 | ━━━━━━━━ 7.5/7.5 MB 69.8 MB/s eta 0:00:00
75 | Downloading torch-2.7.1-cp312-cp312-manylinux_2_28_x86_64.whl (821.0 MB)
76 | ━━━━━━━━ 821.0/821.0 MB 36.3 MB/s eta 0:00:00
77 | Downloading nvidia_cublas_cu12-12.6.4.1-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (393.1 MB)
78 | ━━━━━━━━ 393.1/393.1 MB 84.4 MB/s eta 0:00:00
79 | Downloading nvidia_cuda_cupti_cu12-12.6.80-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (8.9 MB)
80 | ━━━━━━━━ 8.9/8.9 MB 125.3 MB/s eta 0:00:00
81 | Downloading nvidia_cuda_nvrtc_cu12-12.6.77-py3-none-manylinux2014_x86_64.whl (23.7 MB)
82 | ━━━━━━━━ 23.7/23.7 MB 141.0 MB/s eta 0:00:00
83 | Downloading nvidia_cuda_runtime_cu12-12.6.77-py3-none-manylinux2014_x86_64.manylinux_2_17_x86_64.whl (897 kB)
84 | ━━━━━━━━ 897.7/897.7 kB 75.5 MB/s eta 0:00:00
85 | Downloading nvidia_cudnn_cu12-9.5.1.17-py3-none-manylinux_2_28_x86_64.whl (571.0 MB)
```

# Юху, практика!



# Юху, практика!

Namespace: airflow ▾

### Pods

[Создать Pod](#)

Фильтр	Имя	Статус	Готовность	Перезапуски	Владелец	Память	CPU	Дата создания	⋮
	airflow-postgresql-0	Запущен	1/1	0	SS airflow-postgresql	67,6 MiB	0,016 ядра	11 июня 2025 г., 18:41	⋮
	airflow-scheduler-66dfc7bbdf-64jz8	Запущен	3/3	0	RS airflow-scheduler-66dfc7bbdf	1 683,2 MiB	0,126 ядра	11 июня 2025 г., 21:35	⋮
	airflow-statsd-6bddbbb65b-g5qns	Запущен	1/1	0	RS airflow-statsd-6bddbbb65b	24,2 MiB	0,003 ядра	11 июня 2025 г., 18:41	⋮
	airflow-triggerer-0	Запущен	2/2	0	SS airflow-triggerer	1 342,9 MiB	0,055 ядра	11 июня 2025 г., 21:35	⋮
	airflow-webserver-66b7f5d584-hl8nt	Запущен	1/1	0	RS airflow-webserver-66b7f5d584	2 130,7 MiB	0,002 ядра	11 июня 2025 г., 21:35	⋮
	cifar100-classification-train-model-992gx6mk	Запущен	1/1	0	Владелец не указан	-	-	16 июня 2025 г., 12:16	⋮
	train-model-vrpwr2cm	Запущен	1/1	0	Владелец не указан	-	-	16 июня 2025 г., 12:17	⋮

# Юху, практика!

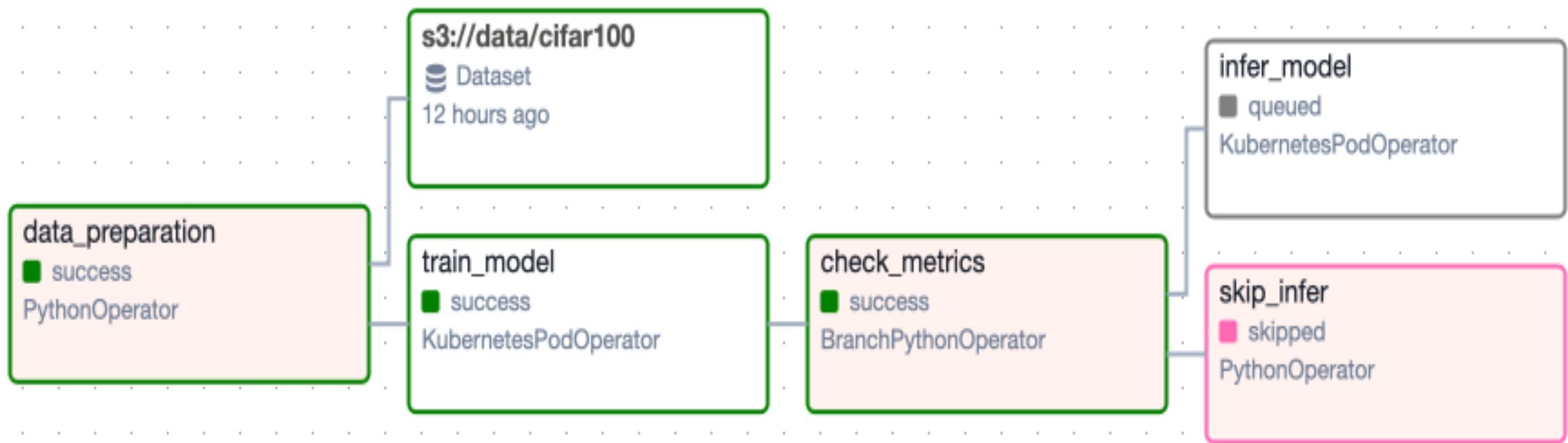
Namespace: airflow ▾

### Pods

Создать Pod

Фильтр	Имя	Поиск по имени...	/	⋮				
Имя	Статус	Готовность	Перезапуски	Владелец	Память	CPU	Дата создания	⋮
 airflow-postgresql-0	 Запущен	1/1	0	 airflow-postgresql	71,7 MiB	0,018 ядра	11 июня 2025 г., 18:41	⋮
 airflow-scheduler-66dfc7bbdf-64jz8	 Запущен	3/3	0	 airflow-scheduler-66dfc7bbdf	1 680,8 MiB	0,131 ядра	11 июня 2025 г., 21:35	⋮
 airflow-statsd-6bddbbb65b-g5qns	 Запущен	1/1	0	 airflow-statsd-6bddbbb65b	23,6 MiB	0,003 ядра	11 июня 2025 г., 18:41	⋮
 airflow-triggerer-0	 Запущен	2/2	0	 airflow-triggerer	1 340,3 MiB	0,053 ядра	11 июня 2025 г., 21:35	⋮
 airflow-webserver-66b7f5d584-hl8nt	 Запущен	1/1	0	 airflow-webserver-66b7f5d584	1 717,8 MiB	0,003 ядра	11 июня 2025 г., 21:35	⋮
 cifar100-classification-check-metrics-q8k4vcov	 Запущен	1/1	0	Владелец не указан	-	-	16 июня 2025 г., 11:53	⋮

# Юху, практика!



# Юху, практика!

## Манифест InferenceService

Дополнительно нужно создать две сущности:

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

```
1  apiVersion: serving.kserve.io/v1beta1
2  kind: InferenceService
3  metadata:
4    name: resnet18
5    namespace: ml-inference
6    annotations:
7      serving.kserve.io/s3-endpoint: "minio.svc.cluster.local"
8      serving.kserve.io/s3-credentials-secret-name: "s3-cred"
9  spec:
10    predictor:
11      serviceAccountName: resnet18-sa
12      triton:
13        runtimeVersion: "23.10-py3"
14        storageUri: "s3://data/models"
15        resources:
16          requests:
17            cpu: "500m"
18            memory: "1Gi"
19            nvidia.com/gpu: 1
20          limits:
21            cpu: "1"
22            memory: "2Gi"
23            nvidia.com/gpu: 1
24    nodeSelector:
25      nvidia.com/gpu.present: "true"
26    tolerations:
27      - key: "nvidia.com/gpu"
28        operator: "Exists"
29        effect: "NoSchedule"
30
```

# Юху, практика!

Namespace: airflow ▾

Pods

Создать Pod

Фильтр	Имя	Статус	Готовность	Перезапуски	Владелец	Память	CPU	Дата создания	⋮
	airflow-postgresql-0	Запущен	1/1	0	airflow-postgresql	72,0 MiB	0,018 ядра	11 июня 2025 г., 18:41	⋮
	airflow-scheduler-66dfc7bbdf-64jz8	Запущен	3/3	0	airflow-scheduler-66dfc7bbdf	1 680,7 MiB	0,139 ядра	11 июня 2025 г., 21:35	⋮
	airflow-statsd-6bdddbb65b-g5qns	Запущен	1/1	0	airflow-statsd-6bdddbb65b	23,7 MiB	0,003 ядра	11 июня 2025 г., 18:41	⋮
	airflow-triggerer-0	Запущен	2/2	0	airflow-triggerer	1 340,4 MiB	0,053 ядра	11 июня 2025 г., 21:35	⋮
	airflow-webserver-66b7f5d584-hl8nt	Запущен	1/1	0	airflow-webserver-66b7f5d584	1 719,6 MiB	0,003 ядра	11 июня 2025 г., 21:35	⋮
	cifar100-classification-infer-model-pjs1hdbi	Запущен	1/1	0	Владелец не указан	-	-	16 июня 2025 г., 11:54	⋮
	gradio-ui-mw0jfasw	Запущен	1/1	0	Владелец не указан	-	-	16 июня 2025 г., 11:55	⋮

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

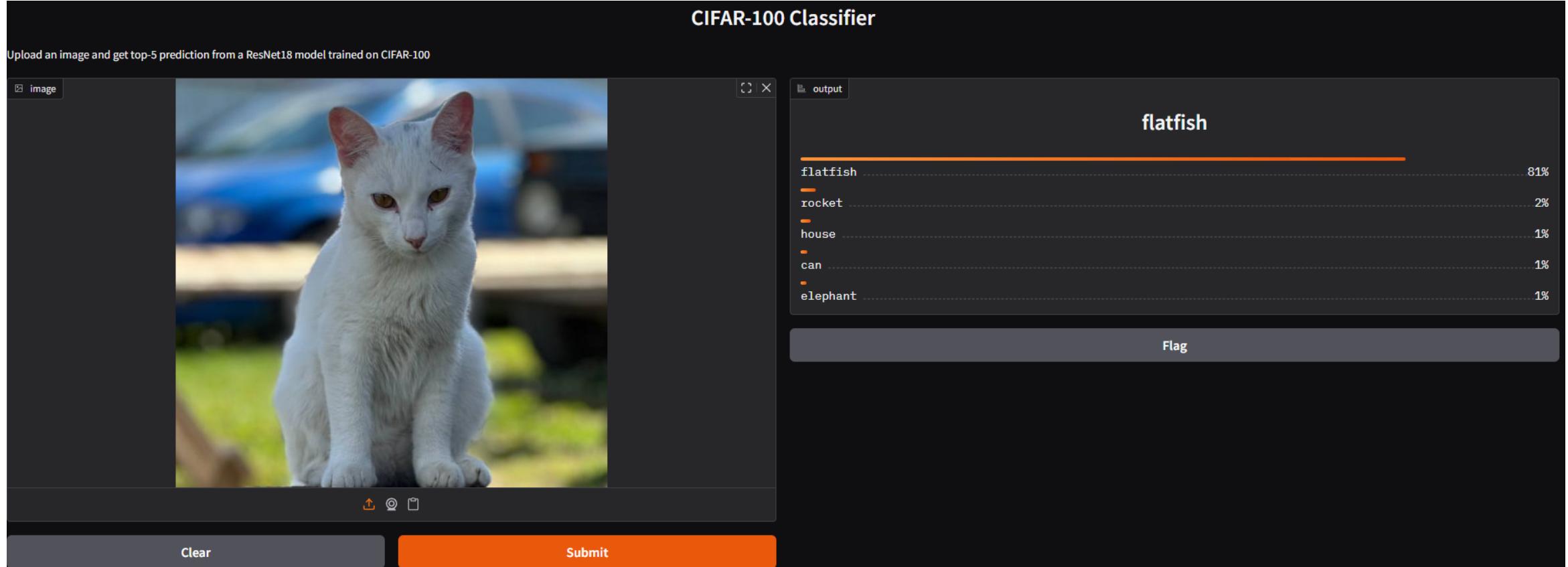
# Юху, практика!

CIFAR-100 Classifier

Upload an image and get top-5 prediction from a ResNet18 model trained on CIFAR-100



# Юху, практика!



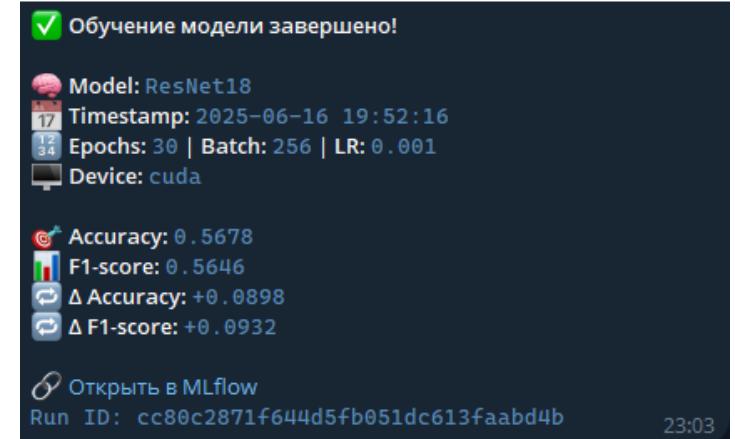
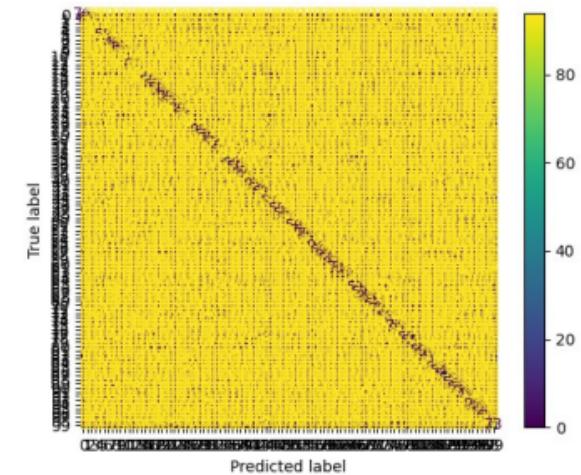
# Юху, практика!

Metrics (6)

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)

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



# Почему песочница «идеальная»?

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

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

Time-to-production ↓ в 2–3 раза

Время обнаружения инцидента ↓  
с часов до минут

Онбординг инженера ↓ с месяцев  
до пары дней

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

99% экспериментов воспроизводимы  
через MLflow

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

NAMi, 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/](https://habr.com/ru/companies/orion_soft/articles/940592/)

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



# Выводы

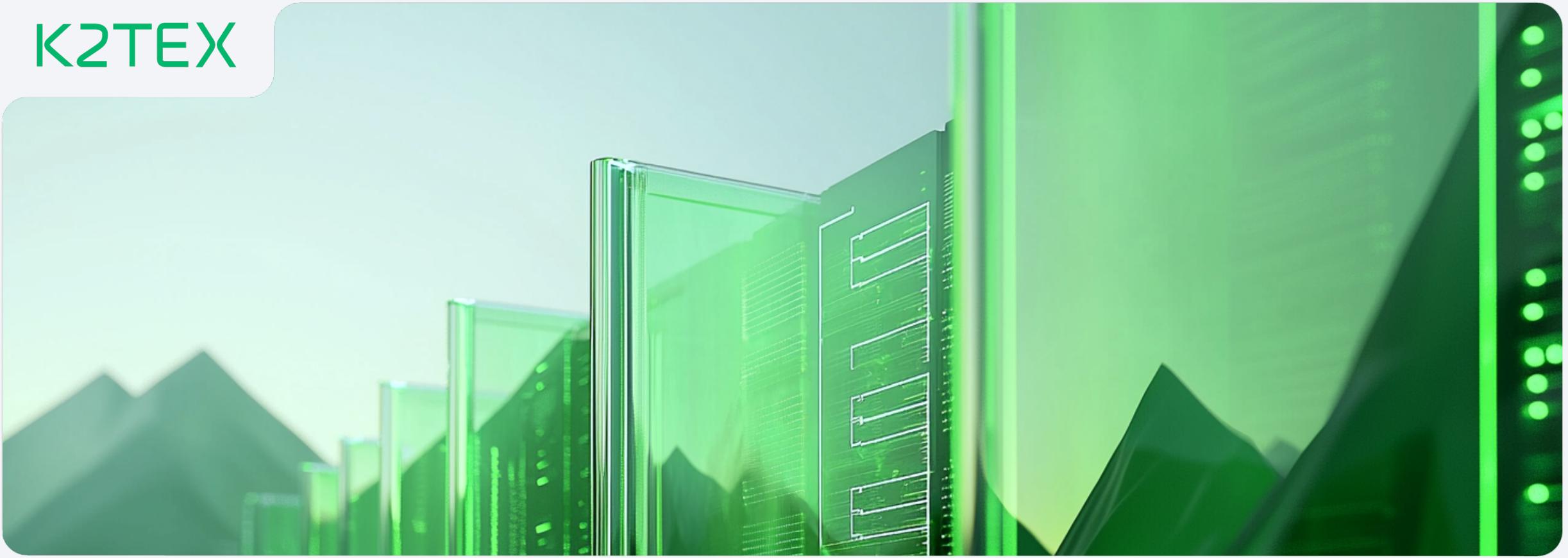
Платформа является кросс-доменной,  
подходит для любой отрасли

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



## Даниил Салман

Tech Lead по практике  
контейнеризации K2 Tech

tg: @DV\_Salman

Будьте в курсе новостей  
K2 TEX

k2.tech  
[info@k2.tech](mailto:info@k2.tech)  
+7 (495) 797-85-84

