



# ML-платформа T-Банка

Что действительно ценят пользователи?





- 5 лет в ML-платформах
- Sber AI
- AI-центр Т-Банка

**Петр Валинг** | Тимлид

@ p.valing@tbank.ru

▶ pvaling

# План



## Кто мы

Как все начиналось и куда пришли



## Топ-фича

Пользователям на самом деле важно



## Тулинг

Чего прикольного понаписали



## Нежданчики

Стоит подумать об этом сразу



## А оно надо?

Стоит ли заморачиваться?

# Наши вводные. Кто мы?

- Т-Банк — цифровая экосистема
- Мобильный оператор / Страхование / Университет / ...
- Много ML/DL специалистов
- Много разнообразных данных
- Разные ML/AI задачи:
  - LLM / TTS / STT / Recsys / OCR / NLP / Research / Classic ML /...
  - Batch / Realtime Inference

# Наш путь

Появление ключевых  
компонентов

Long time ago...

SSH-серваки

2021

Jobs API, Slurm, WandB, Jupyter

2022

Kubernetes и Airflow

2023

Serving Platform, активный рост

2024

WandB -> ClearML, Distributed

2025

Model Registry, LLMaaS

# О платформе. Масштабы



## Нагрузка

- ~1k MAU
- 30k задач в день



## Инфра

- Несколько ДЦ в контуре
- Несколько Cloud
- Сотни серверов
- Разные поколения GPU
- InfiniBand
- Хранилки



## Сервисы

- Compute Jobs
- Jupyter
- Experiment tracking
- Pipelines
- Serving
- LLMaaS
- Model Registry
- ...

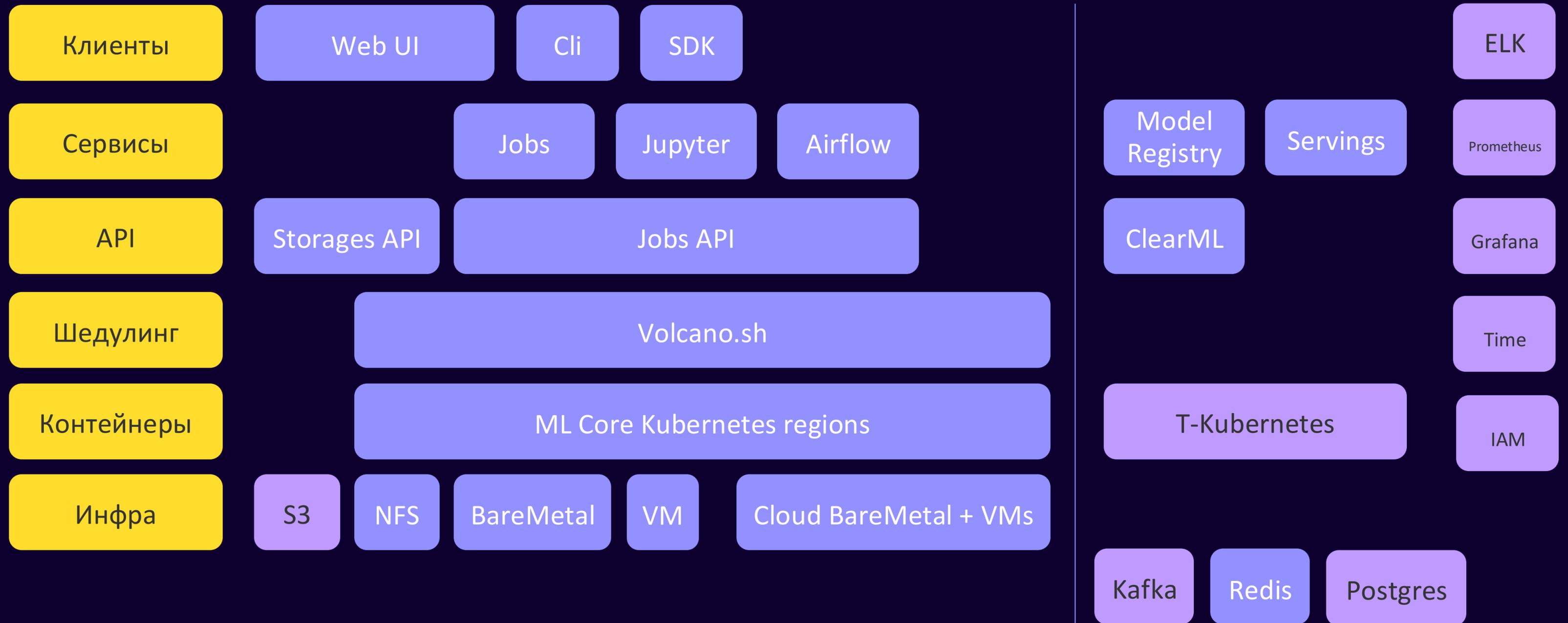


## Команда

- 6+ mid-size команд

# О платформе

По слоям



# Топ фича для пользователей

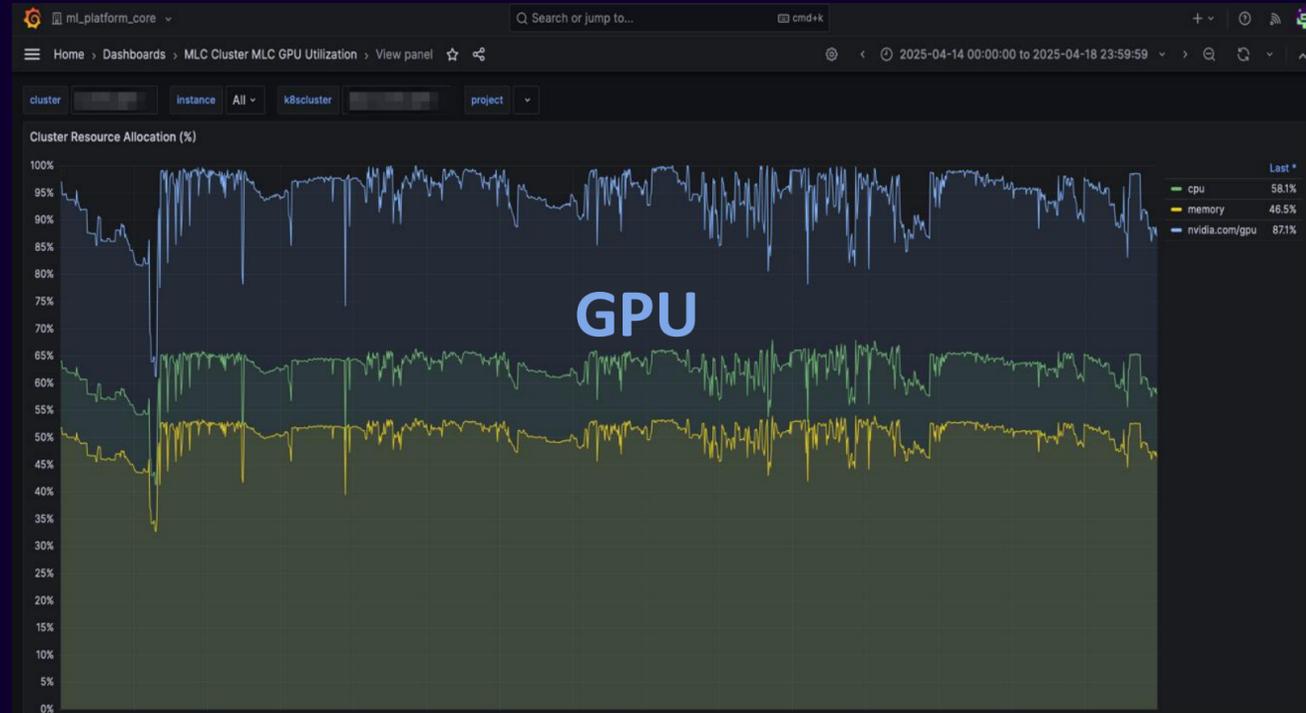
**«Хочу, чтобы никогда  
не надо было ждать ресурсов»**

# Утилизация ресурсов

Метрики из крупных регионов

## Elite GPU

100%



0%

Неделя

## General Purpose GPU



Неделя

# Доступность ресурсов

## Шедулинг задач

- Плотнo упаковываем задачи на нoды
- Data-affinity
- Image-affinity



## Квoты

- Гарантирoванные квoты и лимиты
- Вытесняемые задачи
- Приоритеты и т.д.

## Размеры конфигураций

Фиксированные пропорции CPU/GPU/RAM в каждом регионе (10+ вариантов)

## Минимизация простoя

- Data Load-API
- Кеширование в регионах
- Холодные и горячие кеши

## Тoп-железо

- 8xGPU сервера
- Infiniband / NVMe / Быстрые NFS
- Benchmarks

## Процессы

- Контроль утилизации
- Прогнозирование потребления
- Перераспределение
- Data-driven аналитика

# Доступность ресурсов

Шедулинг задач



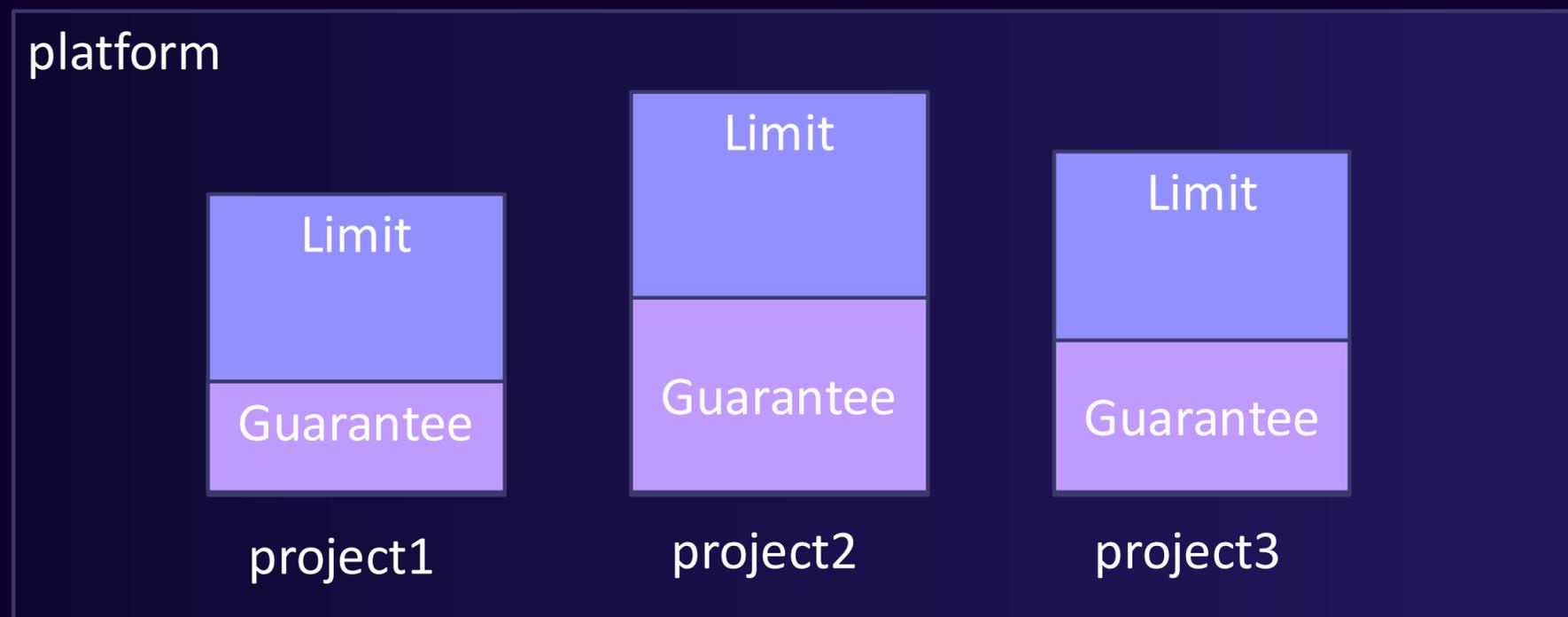
- Volcano.sh
- Плотнo упаковываем задачи
- Data-affinity
- Image-affinity

# Доступность ресурсов

Квоты и лимиты

## Проектные квоты

- Гарантированные: сколько ресурсов зарезервировано под проект
- Лимитирующие: ограничение сверху на количество запрашиваемых ресурсов

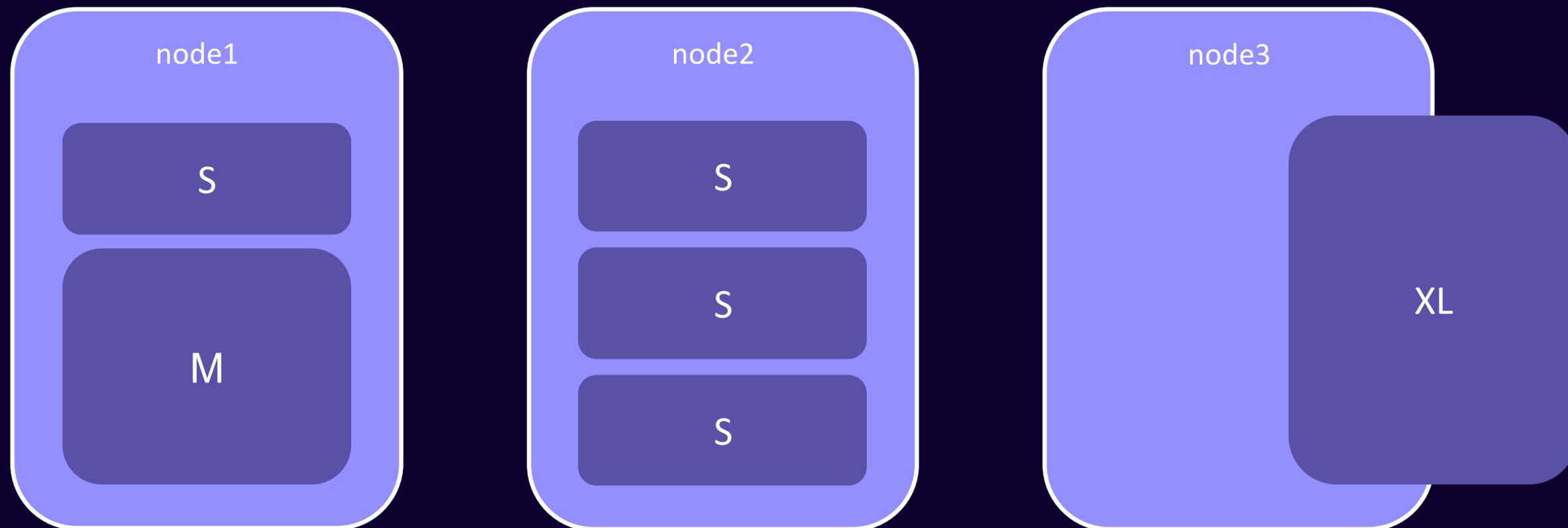


- Вытесняемые задачи
- Подпроекты

# Доступность ресурсов

Размеры конфигураций / flavors

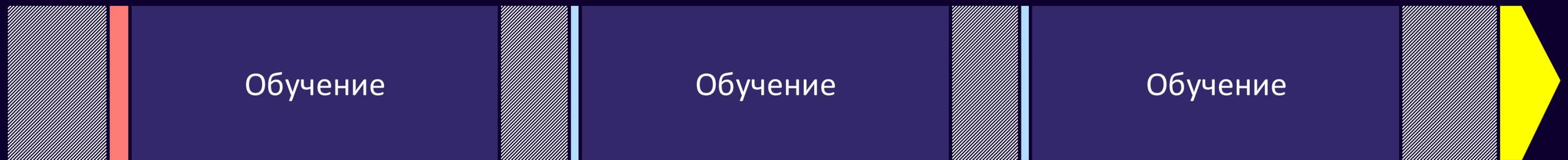
- Фиксированные пропорции CPU/GPU/RAM в каждом регионе (10+ вариантов)



# Доступность ресурсов

Минимизация простоя

## Цикл тренировки модели



Загрузка данных  
и модели в  
горячее  
хранилище  
региона



Отправка чекпоинта  
в горячее  
хранилище региона



Отправка чекпоинта  
в горячее хранилище  
региона



Сохранение артефактов  
с data-api в тенатный  
сторадж

# Доступность ресурсов

Минимизация простоя

- Data Load-API
- Кеширование в регионах
- Холодные и горячие кешы
- Быстрый рестарт задач -> Profit

# Доступность ресурсов

Железо

- 8xGPU сервера
- InfiniBand / NMVE / Быстрые NFS
- Benchmarks

CPUFreq governor

NUMA topo

# Доступность ресурсов

Процессы

- Контроль утилизации
- Прогнозирование потребления
- Перераспределение
- Data-driven аналитика

# Доступность ресурсов

## Шедулинг задач

- Плотнo упаковываем задачи на нoды
- Data-affinity
- Image-affinity

## Квoты

- Гарантирoванные квoты и лимиты
- Вытесняемые задачи
- Приоритеты и т.д.

## Размеры конфигураций

Фиксированные пропорции CPU/GPU/RAM в каждом регионе (10+ вариантов)

## Минимизация простоя

- Data Load-API
- Кеширование в регионах
- Холодные и горячие кеши

## Тoп-железо

- 8xGPU сервера
- Infiniband / NVME / Быстрые NFS
- Benchmarks

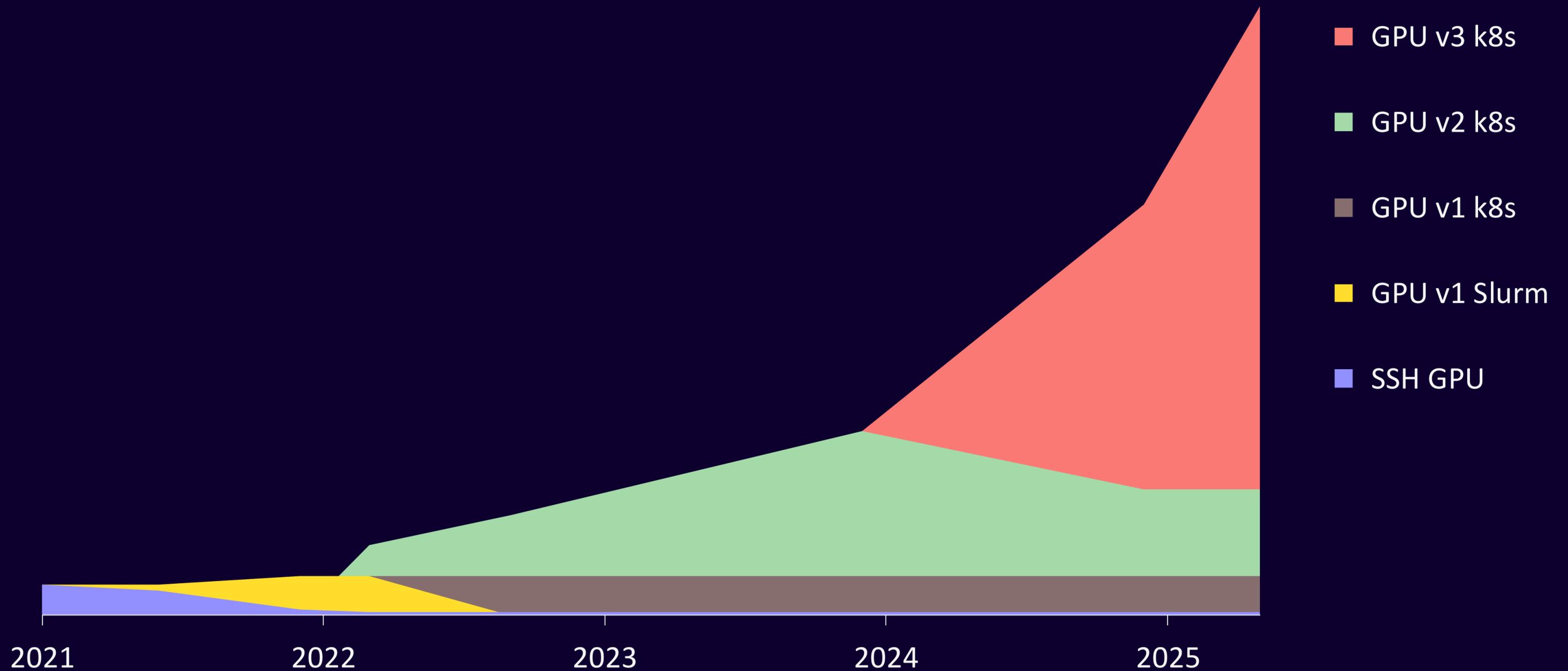
## Процессы

- Контроль утилизации
- Прогнозирование потребления
- Перераспределение
- Data-driven аналитика

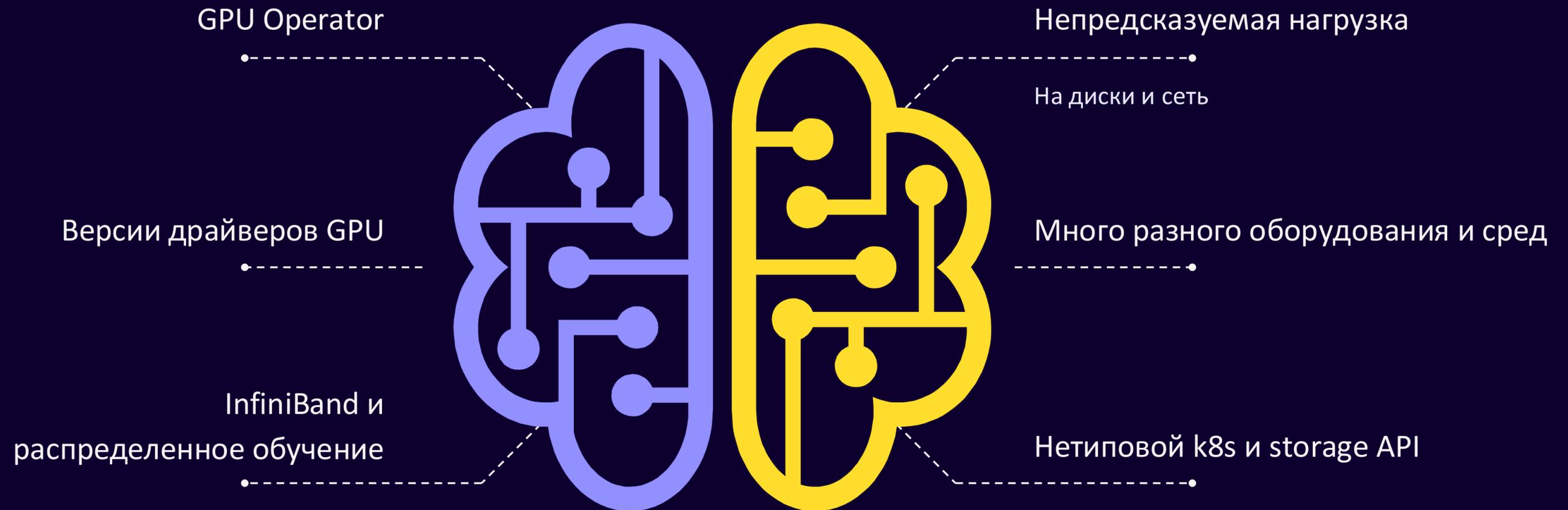
# Кто должен поддерживать ML инфраструктуру?



# Динамика емкости платформы обучения (GPU count)

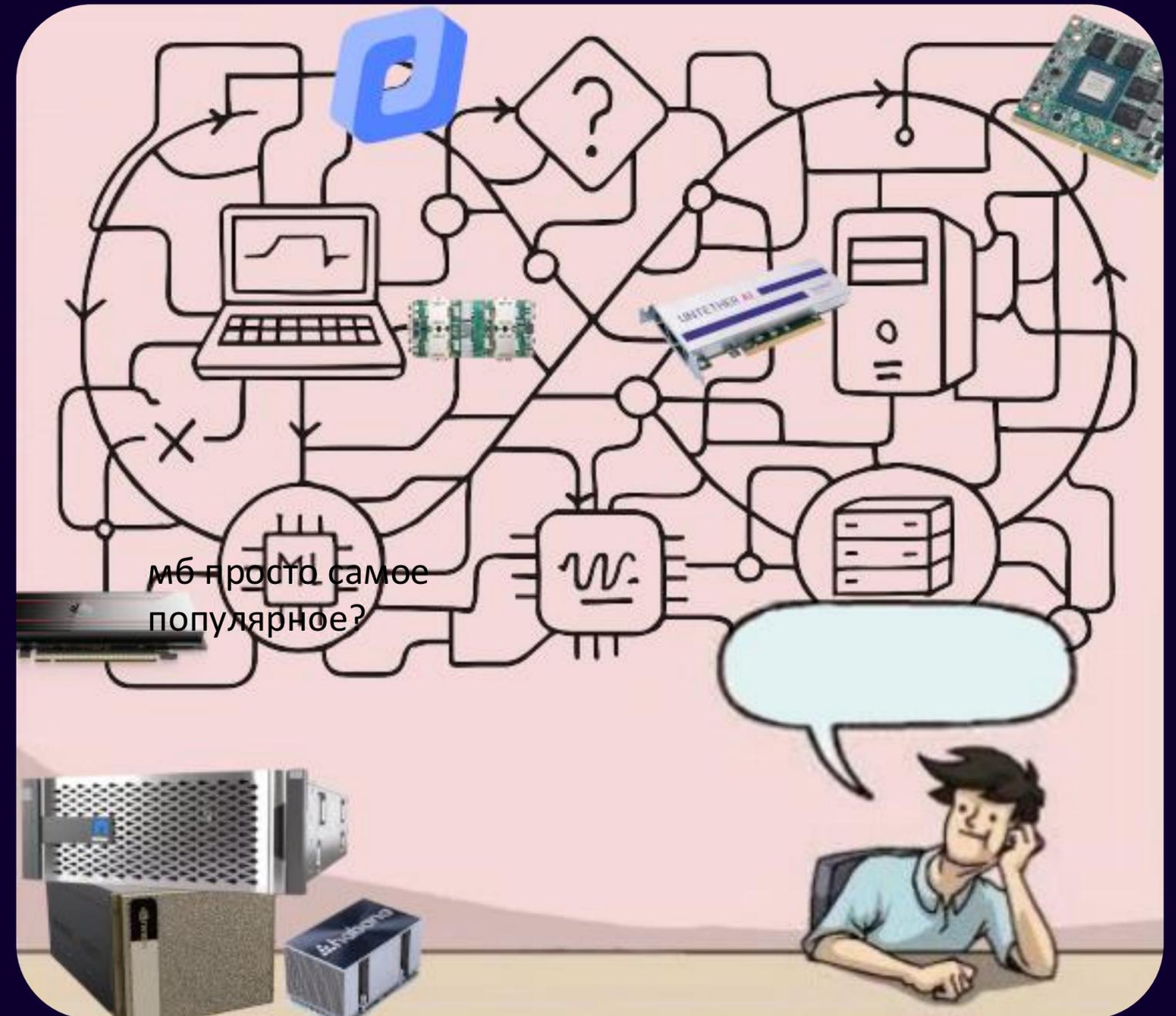


# Высокая когнитивная сложность



# Подбор оптимальных аппаратных ресурсов

- TDP
- Cost
- Benchmarks
- Availability
- Out-scalability
- Cost
- Throughput
- Mem bandwidth
- FP vs Int
- ...



# Отдельная ML Infra Team

ML требует экспериментов на уровне инфраструктуры, что не всегда совместимо с core-business задачами



## Частые перезагрузки

Обновление драйверов и тюнинг производительности



## Эксперименты с сетью

Cilium, DNS, InfiniBand, изоляция



## Использование локальных и сетевых хранилищ

Кеши, логистика артефактов, huggingface-proxy, ...



## Кастомный шедулинг

Требует частого обновления и экспериментов

**Тулинг**

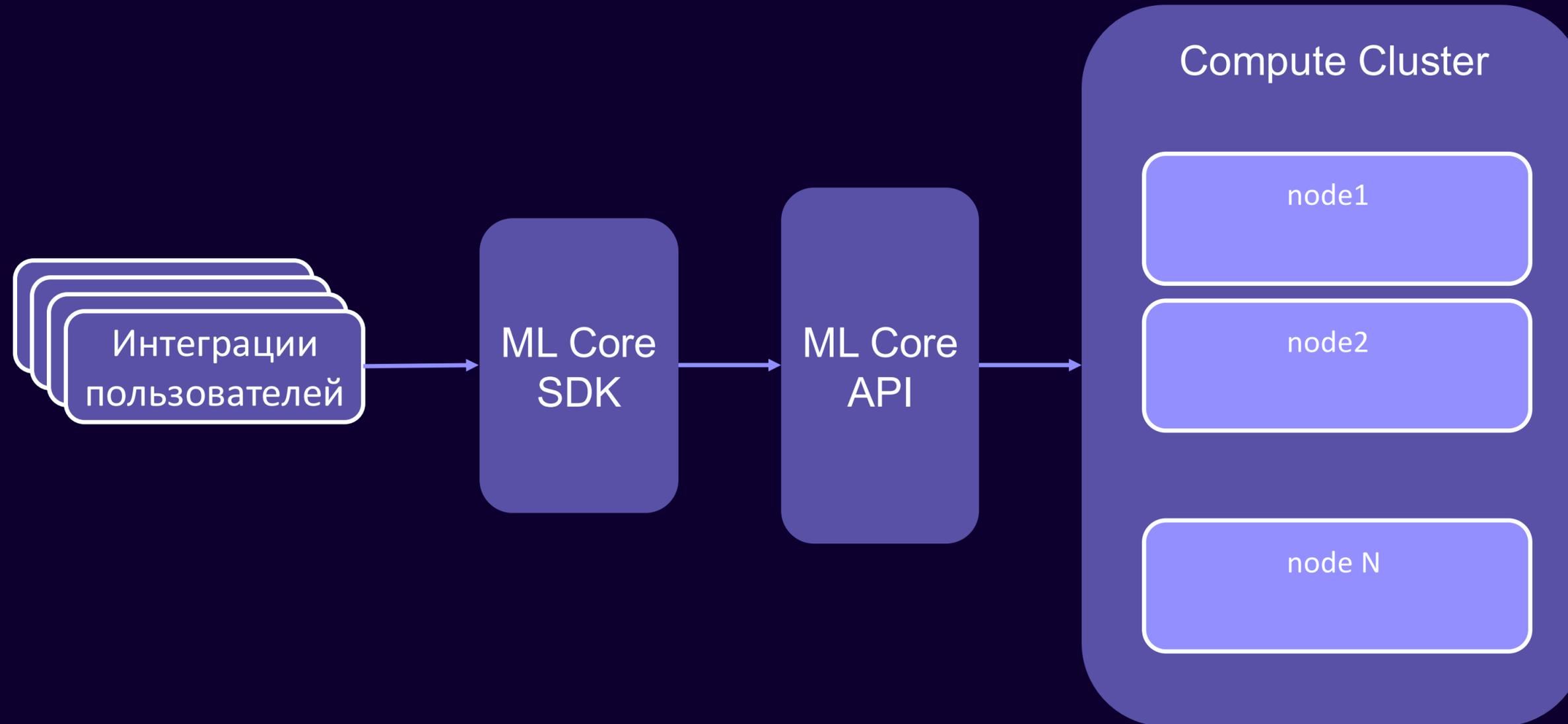
# Тулинг. Compute Jobs



- Собственный стабильный http/grpc API
- Python / Go SDK для интеграций
- Каналы взаимодействия: консольный клиент, веб UI, remote IDE
- Низкий порог входа
- Self-service ops (сетевые доступы, квоты, мониторинг, логи и т.д.)
- Воспроизводимость из коробки

# Тулинг. Compute Jobs

Интеграций будет много



# Тулинг. Compute Jobs



- Собственный стабильный http/grpc API
- Python / Go SDK для интеграций
- Каналы взаимодействия: консольный клиент, веб UI, remote IDE
- Низкий порог входа
- Self-service ops (сетевые доступы, квоты, мониторинг, логи и т.д.)
- Воспроизводимость из коробки

# Тулинг. Compute Jobs

UI для задач

The screenshot displays the ML Core web interface for managing compute jobs. At the top, there is a navigation bar with the ML Core logo and menu items: Ноутбуки, Задачи, Хранилища, Исследования, and Документация. Below the navigation, the page title is 'Задачи'.

### Состояние проекта

Summary statistics for the project:

- 39 (grey)
- 38 (yellow)
- 1 (grey)

**Расход квоты по пользователям**

| Имя          | CPU | GPU |
|--------------|-----|-----|
| sa-ge-mlcore | 232 | 11  |
| mlcore       | 96  | 8   |
| mlcore       | 36  | -   |

**CPU: 316** > Доступно

**Квоты** ? Лимит: 680  
Занято: 364

**GPU: 16** > Доступно

**Квоты** ? Лимит: 35  
Занято: 19

Buttons: Все запуски, Поиск задач по атрибутам

Filters: **Работают** 34, Ожидают 4

| Название                | Ресурсы                    | CPU        | GPU        | RAM        | Создано                         | Until deadline | Тип           |
|-------------------------|----------------------------|------------|------------|------------|---------------------------------|----------------|---------------|
| n-train-g6d8y6          | 12 CPU, 1 GPU, 128 GiB RAM | ██████████ | ██████████ | ██████████ | sa-mlcore<br>2 May 2025 - 20:26 | час            | Airflow On-De |
| nsformer-inference...   | 8 CPU, 64 GiB RAM          | ██████████ | ██████████ | ██████████ | sa-mlcore<br>2 May 2025 - 21:11 | 7 дней         | On-De         |
| i2i-recent-inference... | 8 CPU, 160 GiB RAM         | ██████████ | ██████████ | ██████████ | sa-mlcore<br>2 May 2025 - 20:57 | 19 минут       | On-De         |
| nsformer-inference...   | 8 CPU, 64 GiB RAM          | ██████████ | ██████████ | ██████████ | sa-mlcore<br>2 May 2025 - 20:53 | 7 дней         | On-De         |
| nsformer-inference...   | 8 CPU, 64 GiB RAM          | ██████████ | ██████████ | ██████████ | sa-mlcore<br>2 May 2025 - 20:53 | 7 дней         | Airflow On-De |
| ence-s3-encode-us...    | 12 CPU, 1 GPU, 128 GiB RAM | ██████████ | ██████████ | ██████████ | sa-mlcore<br>2 May 2025 - 20:49 | 12 часов       | Airflow On-De |

# Тулинг. Compute Jobs

## UI для задач

The screenshot displays the ML Core interface for a specific compute job. The top navigation bar includes 'ML Core' and various menu items like 'Ноутбуки', 'Задачи', 'Хранилища', 'Исследования', and 'Документация'. The breadcrumb trail shows 'Задачи > История запусков'. The job title is 'periment\_2-d2-0-98r5xt' and its status is 'Работает' (Running).

**Job Metadata:**

- Регион: [Redacted]
- TTL: 6 d 23 hr
- Приоритет: [Yellow bar]
- Фаза: Started
- Автор: @a [Redacted]
- Дата создания: 28 Apr 2025 - 18:45
- Проект: [Redacted]
- Мониторинг: Grafana
- Clear ML Tasks: [Redacted]
- Job: [Redacted]
- Vault Config: {"authRole": [Redacted]}
- Security Class: DEFAULT\_JOB
- Вложения: Здесь пока ничего нет
- Артефакты: Здесь пока ничего нет
- Образы: К сведениям [Info icon]
- qos: OnDemand
- Логи Airflow: Ссылка
- Лейблы: -

**Логи </>** (Log viewer):

```
2025-04-28 17:58:23.6267Z | de | train
2025-04-28 17:58:23.6673Z | de | train
2025-04-28 17:58:23.7452Z | de | train
2025-04-28 17:58:23.7474Z | cr | train
2025-04-28 17:58:23.7865Z | ac | train
2025-04-28 17:58:23.8267Z | ac | train
2025-04-28 17:58:23.8674Z | ac | train
-----
2025-04-28 17:58:23.9469Z |   | Trainable params
2025-04-28 17:58:23.9469Z |   | Non-trainable params
2025-04-28 17:58:23.9871Z |   | Total params
2025-04-28 17:58:24.0273Z |   | Total estimated model params size (MB)
```

**Контейнеры** (Containers):

- periment\_2-d2-0-98r5xt [Info icon]

**Terminal:**

```
root@k8smr-...-ex-3002849303-task-0:/workspaces#
```

# Тулинг. Compute Jobs

## Консольный клиент ML Core

```
p.valing — p.valing@macbook-Q7PXQVF6L5 — ~ — zsh — 128x47
→ ~ mlc job --help
Usage: mlc job <command>

Submit, control and describe running or completed jobs

Commands:
  ls                List all jobs in the currently selected project

  get               Get extensive job description in YAML format
  <JOB_NAME>       Name of the job to describe

  logs             Stream or search running or completed job logs
  <JOB_NAME>       Name of the job to print logs for

  submit           Submit a new job for execution and optionally tail its logs
  [<IMAGE>]        Docker image to run
  [<args> ...]      Arguments to pass to the entrypoint (default specified in Dockerfile or overridden via --entrypoint)

  lint             Lint job preset file or command line
  [<IMAGE>]        Docker image to run
  [<args> ...]      Arguments to pass to the entrypoint (default specified in Dockerfile or overridden via --entrypoint)

  test            Submit a new job in gpu-test environment
  [<IMAGE>]        Docker image to run
  [<args> ...]      Arguments to pass to the entrypoint (default specified in Dockerfile or overridden via --entrypoint)

  cancel          Cancel running or pending job
  <JOB_NAME>       Name of the job to cancel

  preempt         Manually preempt a running job
  <JOB_NAME>       Name of the job to manually preempt

  rerun           Rerun a job
  <JOB_NAME>       Name of the job to rerun

  download        Download job
  <type>          Specify one of resource type: artifacts, attachments, logs, presets
  <JOB_NAME>       Name of the job

  port-forward    Forward port to job
  <JOB_NAME>       Name of the job to print logs for

  exec            Execute command in job
  <JOB_NAME>       Name of the job to print logs for

Flags:
  -h, --help      Show context-sensitive help.
```

```
mlc_tutorial — p.valing@macbook-Q7PXQVF6L5 — ../mlc_...
[→ mlc_tutorial ls
main.py
→ mlc_tutorial █
```

# Тулинг. Compute Jobs



- Собственный стабильный http/grpc API
- Python / Go SDK для интеграций
- Каналы взаимодействия: консольный клиент, веб UI, remote IDE
- Низкий порог входа
- Self-service ops (сетевые доступы, квоты, мониторинг, логи и т.д.)
- Воспроизводимость из коробки

# Тулинг. Jupyter



- На базе технологии Jobs
- Подключение данных и секретов
- Когда выключать ноутбуки?
- Сборка образов
- Расширяемость: observability / сниппеты
- AI-ассистенты

A screenshot of the ML Core JupyterLab interface. The top navigation bar includes the ML Core logo, a menu with "Ноутбуки", "Задачи", "Хранилища", "Исследования", and "Документация", and a user profile "synopsis". Below the navigation bar, the "Ноутбуки" section is active, showing a search bar and a list of notebooks. The list has columns for "Название", "Статус", "Ресурсы", "Хранилище", "Последний запуск", and "Дата соз". Two notebooks are visible: "synopsis-bot" (status: "Запускается", resources: "1 CPU 2 GiB RAM", storage: "0 MiB / 8 GiB", last run: "2 May 2025 - 19:33") and "gp-test" (status: "Остановлен", resources: "2 CPU 10 GiB RAM", storage: "0 MiB / 8 GiB"). A context menu is open over the "gp-test" notebook, showing options "Редактировать настройки" and "Использовать как шаблон".

| Название     | Статус      | Ресурсы          | Хранилище     | Последний запуск               | Дата соз |
|--------------|-------------|------------------|---------------|--------------------------------|----------|
| synopsis-bot | Запускается | 1 CPU 2 GiB RAM  | 0 MiB / 8 GiB | p.valing<br>2 May 2025 - 19:33 |          |
| gp-test      | Остановлен  | 2 CPU 10 GiB RAM | 0 MiB / 8 GiB |                                |          |

# Тулинг. Jupyter

Сборка образов для ноутбуков

The screenshot shows the JupyterLab interface with a Dockerfile editor open. The editor displays the following Dockerfile content:

```
1 FROM docke
python-3.9
2
3 # Если вам нужно установить какие то pip пакеты вы можете сделать это с
помощью директивы RUN
4 # RUN pip install --no-cache-dir <package-name>
5
6 # Так же вы можете установить необходимые зависимости с помощью пакетного
менеджера.
7 # Пример установки пакета с помощью apt-get:
8 # RUN sudo apt-get update && apt-get install -y <package-name>
9
10 # Если вам нужно во время сборки использовать root пользователя, то вы можете
сделать это с помощью директивы USER.
11 # Но при этом не забудьте сделать возврат в стандартного пользователя mlcore в
образе тетрадки.
12 # Это необходимо, чтобы вы не получали ошибки при запуске тетрадки.
13 # USER root
```

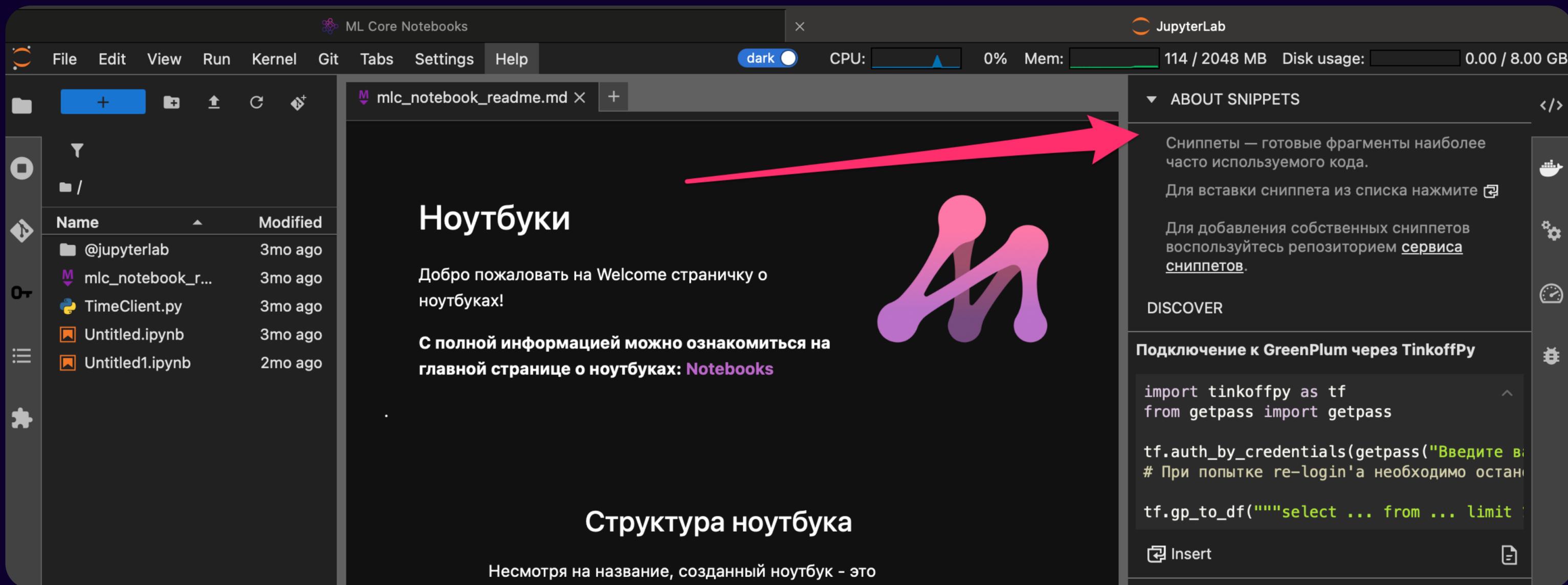
The interface includes a file explorer on the left showing the file structure, a search bar at the top with the text "обр", and a right-hand panel with the following fields:

- Имя образа: time-msg
- Версия образа: 1.0.0
- Зарегистрировать образ в ML Core
- Собрать образ (button)

The top status bar shows system metrics: CPU: 0%, Mem: 114 / 2048 MB, and Disk usage: 0.00 / 8.00 GB.

# Тулинг. Jupyter

Сниппеты кода в ноутбуках



The screenshot shows the JupyterLab interface. The main window displays a notebook titled "Ноутбуки" (Notebooks) with the following content:

## Ноутбуки

Добро пожаловать на Welcome страничку о ноутбуках!

С полной информацией можно ознакомиться на главной странице о ноутбуках: [Notebooks](#)

### Структура ноутбука

Несмотря на название, созданный ноутбук - это

A red arrow points from the main content area to the "ABOUT SNIPPETS" panel in the right sidebar. The sidebar panel contains the following text:

#### ABOUT SNIPPETS

Сниппеты — готовые фрагменты наиболее часто используемого кода.

Для вставки сниппета из списка нажмите 

Для добавления собственных сниппетов воспользуйтесь репозиторием [сервиса сниппетов](#).

#### DISCOVER

##### Подключение к GreenPlum через TinkoffPy

```
import tinkoffpy as tf
from getpass import getpass

tf.auth_by_credentials(getpass("Введите в:
# При попытке re-login'a необходимо остан

tf.gp_to_df("""select ... from ... limit
```

At the bottom of the code block, there is an "Insert" button with a copy icon.

# Тулинг. Jupyter

- На базе технологии Jobs
- Подключение данных и секретов
- Когда выключать ноутбуки?
- Сборка образов
- Расширяемость: `observability` / сниппеты
- AI-ассистенты



# Тулинг. Managed Apache Airflow



- ML-пайплайны - must
- Сила Open Source-бренда
- Сложно освоить, но зато легко начать
- Уже 250+ инстансов
- Нативный доступ к фичам и ресурсам  
ML-платформы / общий тулинг и UX
- Сборка образов
- Маркетплейс полезных операторов

# Тулинг. Managed Apache Airflow

## ML-пайплайны в Airflow

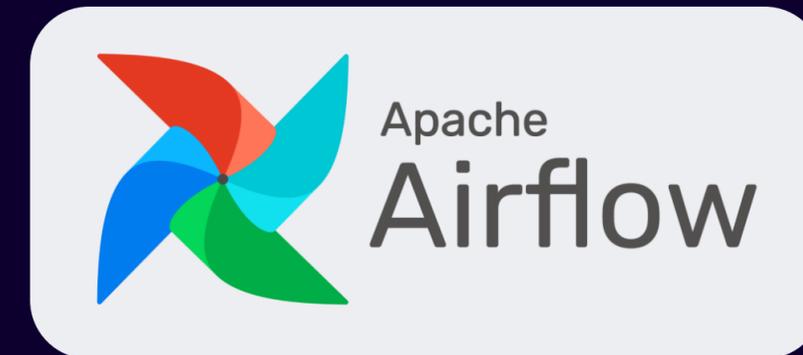
The screenshot displays the Apache Airflow web interface for a DAG named 'gp\_etl\_dag'. The interface includes a top navigation bar with 'Airflow' logo and menu items like 'DAGs', 'Cluster Activity', 'Datasets', 'Security', 'Browse', 'Admin', and 'Docs'. The current time is 18:07 MSK (+03:00). The DAG title is 'DAG: gp\_etl\_dag' with a subtitle 'Пример ETL процесса на Greenplum'. The interface shows a filter for the run date '12/10/2024 10:51:10 AM' and 'All Run Types' and 'All Run States'. A legend at the top lists various task states: deferred, failed, queued, removed, restarting, running, scheduled, shutdown, skipped, success, up\_for\_reschedule, up\_for\_retry, upstream\_failed, and no\_status. The main view shows a DAG run for '2024-12-09, 21:48:53 MSK'. The DAG graph consists of the following tasks:

- load\_initial\_data (success, CSVToGreenplumOperator)
- make\_dataset\_table (success, GreenplumExecuteOperator)
- fetch\_dataset\_from\_gp (success, GreenplumToDataframeOperator)
- make\_python\_processing (success, PythonOperator)
- make\_mlc\_job\_processing (success, MLCoreJobSubmitOperator)
- join\_results (success, PythonOperator)
- upload\_predictions (success, CSVToGreenplumOperator)
- drop\_tables\_from\_gp (success, GreenplumExecuteOperator)

On the left, a Gantt chart shows the duration of each task, with a total duration of 00:37:39. A table below the Gantt chart lists the tasks and their durations:

| Task Name               | Duration |
|-------------------------|----------|
| load_initial_data       | 00:00:00 |
| make_dataset_table      | 00:00:00 |
| fetch_dataset_from_gp   | 00:00:00 |
| make_python_processing  | 00:00:00 |
| make_mlc_job_processing | 00:00:00 |
| join_results            | 00:00:00 |
| upload_predictions      | 00:00:00 |
| drop_tables_from_gp     | 00:00:00 |

# Тулинг. Managed Apache Airflow



- ML-пайплайны - must
- Сила Open Source-бренда
- Сложно освоить, но зато легко начать
- Уже 250+ инстансов
- Нативный доступ к фичам и ресурсам  
ML-платформы / общий тулинг и UX
- Сборка образов
- Маркетплейс полезных операторов

# Тулинг. Managed Apache Airflow

Управление инстансами Airflow

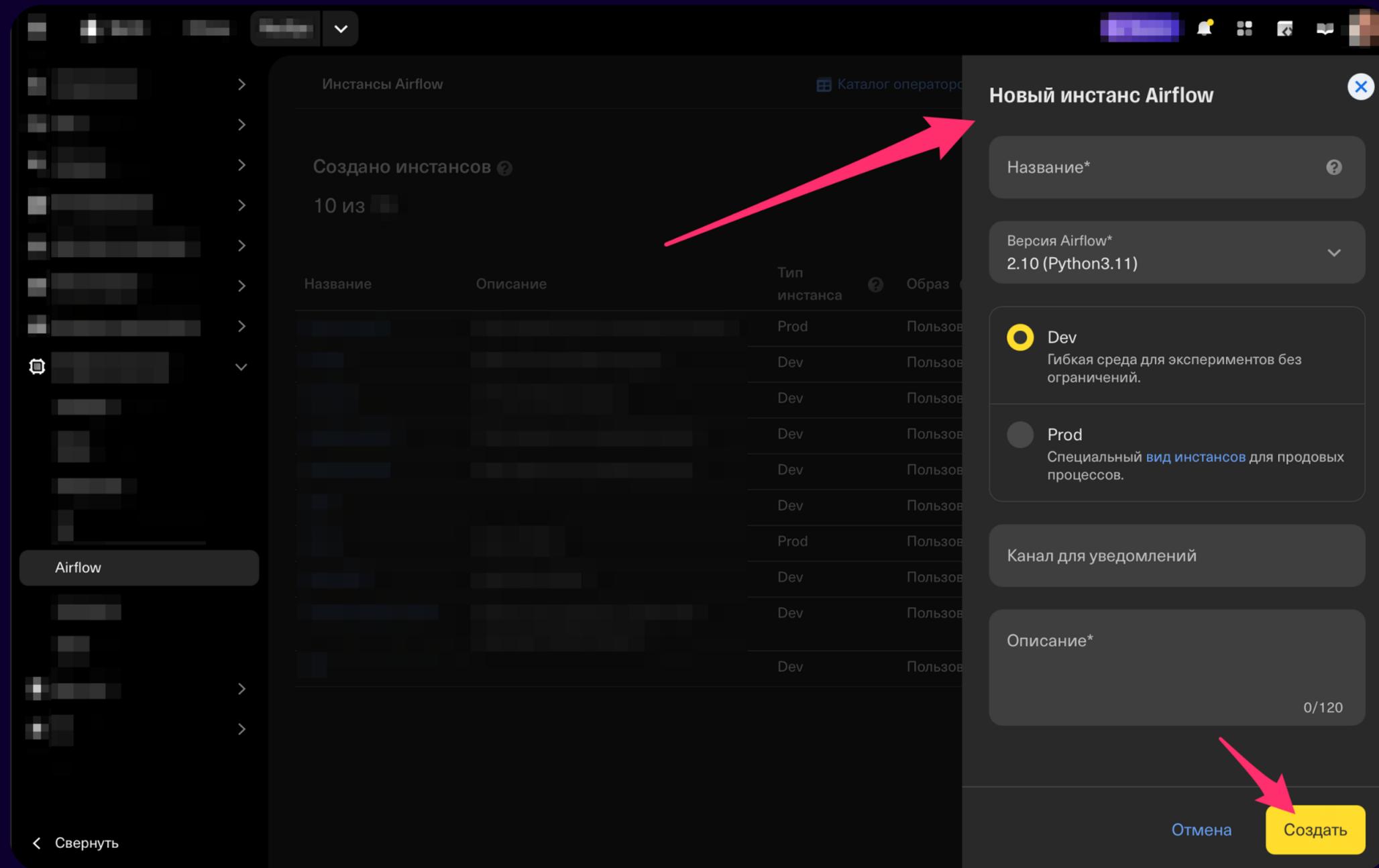
The screenshot displays the 'Инстансы Airflow' (Airflow Instances) management interface. At the top right, there are navigation links for 'Каталог операторов', 'Описание услуг', and a highlighted 'Создать инстанс' (Create Instance) button, which is pointed to by a red arrow. Below the navigation, it indicates 'Создано инстансов 10 из 15' (10 instances created out of 15). The main area contains a table with the following columns: 'Название' (Name), 'Описание' (Description), 'Тип инстанса' (Instance Type), 'Образ' (Image), 'Версия Airflow' (Airflow Version), 'Airflow UI', and 'State'. The table lists 10 instances with various configurations and states.

| Название   | Описание   | Тип инстанса | Образ            | Версия Airflow | Airflow UI | State      |
|------------|------------|--------------|------------------|----------------|------------|------------|
| [Redacted] | [Redacted] | Prod         | Пользовательский | 2.9.2          | Airflow    | Работает   |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.9.2          | Airflow    | Работает   |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.9.2          | Airflow    | Работает   |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.9.2          | -          | Остановлен |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.9.2          | Airflow    | Работает   |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.9.2          | Airflow    | Работает   |
| [Redacted] | [Redacted] | Prod         | Пользовательский | 2.9.2          | Airflow    | Работает   |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.10           | Airflow    | Работает   |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.10           | Airflow    | Работает   |
| [Redacted] | [Redacted] | Dev          | Пользовательский | 2.9.2          | Airflow    | Работает   |

A red arrow on the left side of the interface points to the 'Airflow' menu item in the sidebar.

# Тулинг. Managed Apache Airflow

Создание нового инстанса Airflow



# Тулинг. Managed Apache Airflow



- ML-пайплайны - must
- Сила Open Source-бренда
- Сложно освоить, но зато легко начать
- Уже 250+ инстансов
- Нативный доступ к фичам и ресурсам  
ML-платформы / общий тулинг и UX
- Сборка образов
- Маркетплейс полезных операторов

# Тулинг. Managed Apache Airflow

## Сборка кастомных образов для Airflow

```
mlc airflow images ls -p
```

| ID    | STATE | CFG_ID | CFG_NAME       | CFG_DESC                        | VERSION | BUILD           |
|-------|-------|--------|----------------|---------------------------------|---------|-----------------|
| 15629 | Built | 12032  | nary           | ry-configuration                | 2.9.2   | release-v0.5.28 |
| 15431 | Built | 11841  | nary           | ry-configuration                | 2.9.2   | release-v0.5.28 |
| 15383 | Built | 11798  | nary           | ry-configuration                | 2.9.2   | release-v0.5.28 |
| 15292 | Built | 11710  | nary           | ry-configuration                | 2.9.2   | release-v0.5.28 |
| 15257 | Built | 11684  | ed-test        | -configuration                  | 2.9.2   | release-v0.5.27 |
| 15256 | Built | 11683  | ed-test        | -configuration                  | 2.9.2   | release-v0.5.27 |
| 15226 | Built | 11653  | -legacy-prod-v | Prod Airflow Configuration v3.0 | 2.9.2   | release-v0.5.27 |
| 15219 | Built | 11647  | nary           | ry-configuration                | 2.9.2   | release-v0.5.27 |
| 15218 | Built | 11646  | nary           | ry-configuration                | 2.9.2   | release-v0.5.27 |
| 15217 | Built | 11645  | nary           | ry-configuration                | 2.9.2   | release-v0.5.27 |
| 15216 | Built | 11644  | ed-test        | -configuration                  | 2.9.2   | release-v0.5.27 |
| 15214 | Built | 11642  | nary           | ry-configuration                | 2.9.2   | release-v0.5.27 |
| 15210 | Built | 11639  | nary           | ry-configuration                | 2.9.2   | release-v0.5.27 |

ML Core Airflow Private > airflow\_custom\_image > Pipelines > #63631589

passed Pipeline #63631589 triggered 2 days ago by Service bot

### Startup tests improvement

Pipeline Needs Jobs 8 Tests 0

Group jobs by Stage Job dependencies

| validate      | build | test                 | publish    |
|---------------|-------|----------------------|------------|
| check image   | build | log packages         | update tag |
| info          |       | test airflow startup |            |
| prebuild-reqs |       | test dependencies    |            |

# Тулинг. Managed Apache Airflow



- ML-пайплайны - must
- Сила Open Source-бренда
- Сложно освоить, но зато легко начать
- Уже 250+ инстансов
- Нативный доступ к фичам и ресурсам  
ML-платформы / общий тулинг и UX
- Сборка образов
- Маркетплейс полезных операторов

# Тулинг. Managed Apache Airflow

## Каталог готовых операторов для Airflow

ML Core | Документация | TiMe | Репозиторий | Nestor | Искать везде | Search

- добавляйте в код как зависимость (например, `from airflow_provider_mlcore.operators.<***> import <***>`)
- [blurred]
- [blurred]

### Готовые операторы

| Оператор                                      | Назначение   | Участвующие системы |
|---|--|---------------------|
| <a href="#">MLCoreJobSubmitOperator</a>       | создает и запускает задачу в MLC Jobs  | ML Core             |
| <a href="#">MLCoreJobStatusSensor</a>         | сенсор, следящий за статусом задачи в MLC Jobs                               | ML Core             |
| <a href="#">GreenplumToDataframeOperator</a>  | выполняет запрос в GP через <code>tinkoffpy.gp_to_df</code>                  | Greenplum           |
| <a href="#">GreenplumToCSVOperator</a>        | выполняет запрос в GP через <code>tinkoffpy.gp_to_df</code>                  | Greenplum           |
| <a href="#">CSVToGreenplumOperator</a>        | выполняет запрос в GP через <code>tinkoffpy.csv_to_gp</code>                 | Greenplum           |
| <a href="#">GreenplumExecuteOperator</a>      | выполняет запрос в GP через <code>tinkoffpy.gp_execute</code>                | Greenplum           |
| <a href="#">GreenplumTablesWaitSensor</a>     | ожидает актуальность таблицы в GP через <code>tinkoffpy.tables_status</code> | Greenplum           |
| <a href="#">[blurred]OrderCompletedSensor</a> | возвращает статус заказа [blurred]   | [blurred]           |
| <a href="#">SparkToDFOperator</a>             | выполняет запрос в Spark   | DLH/Spark           |
| <a href="#">SparkCursorOperator</a>           | возвращает курсор для выполнения запросов в Spark                            | DLH/Spark           |

# Тулинг. Managed Apache Airflow



- ML-пайплайны - must
- Сила Open Source-бренда
- Сложно освоить, но зато легко начать
- Уже 250+ инстансов
- Нативный доступ к фичам и ресурсам  
ML-платформы / общий тулинг и UX
- Сборка образов
- Маркетплейс полезных операторов

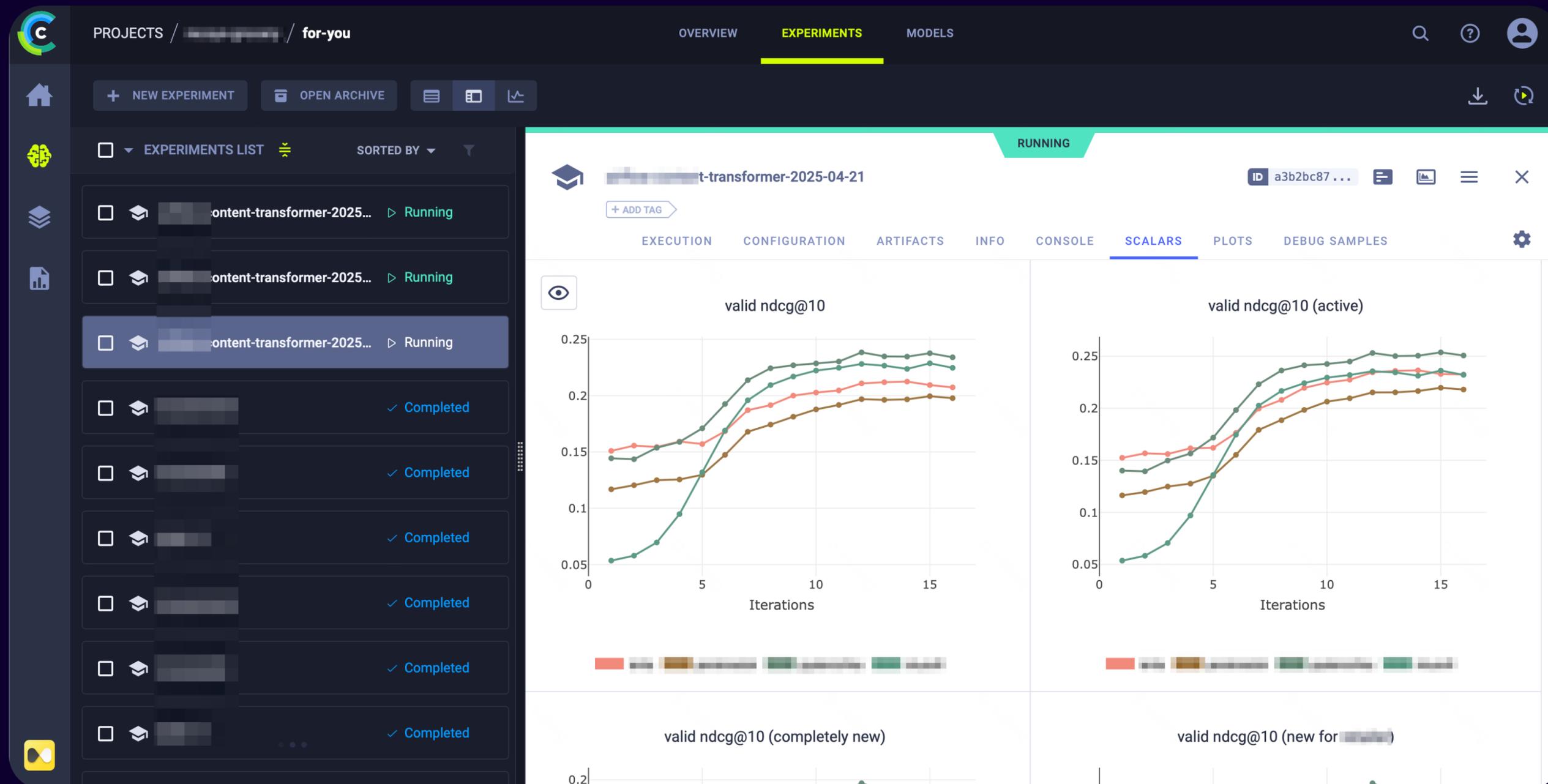
# Тулинг. Трекинг экспериментов / ClearML

- Бесшовная интеграция с Jobs / Jupyter / ModelRegistry / Serving
- Самописная мультитенантность
- +- Кастомизируется
- Держит нагрузку и масштабируется
- Rate limiter и retention (!)
- Тех. независимость / in-house
- Есть приколы... (файлсервер, клиент)



# Тулинг. Трекинг экспериментов / ClearML

UI трекера экспериментов



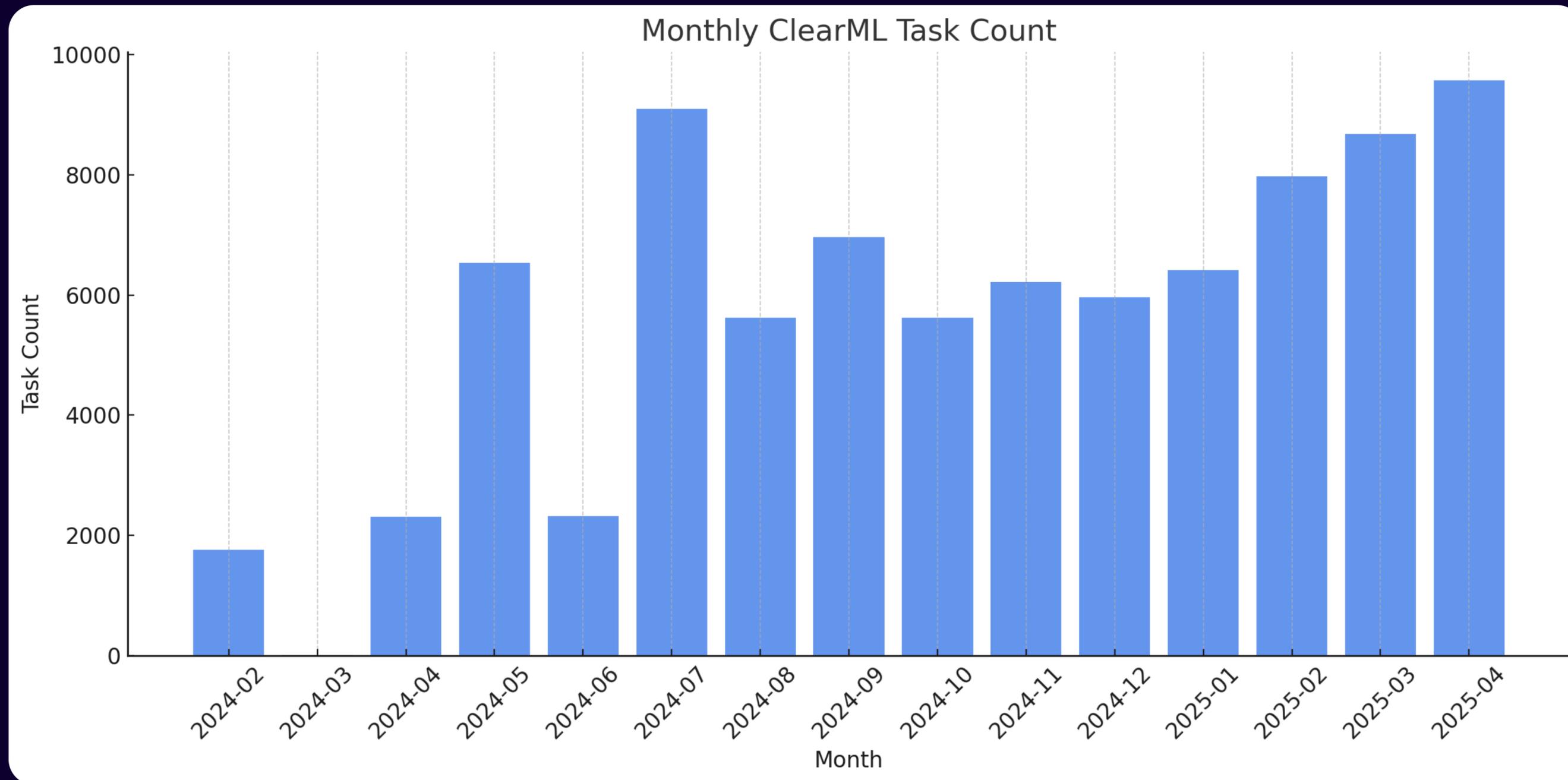
# Тулинг. Трекинг экспериментов / ClearML

- Бесшовная интеграция с Jobs / Jupyter / ModelRegistry / Serving
- Самописная мультитенантность
- +- Кастомизируется
- Держит нагрузку и масштабируется
- Rate limiter и retention (!)
- Тех. независимость / in-house
- Есть приколы... (файлсервер, клиент)



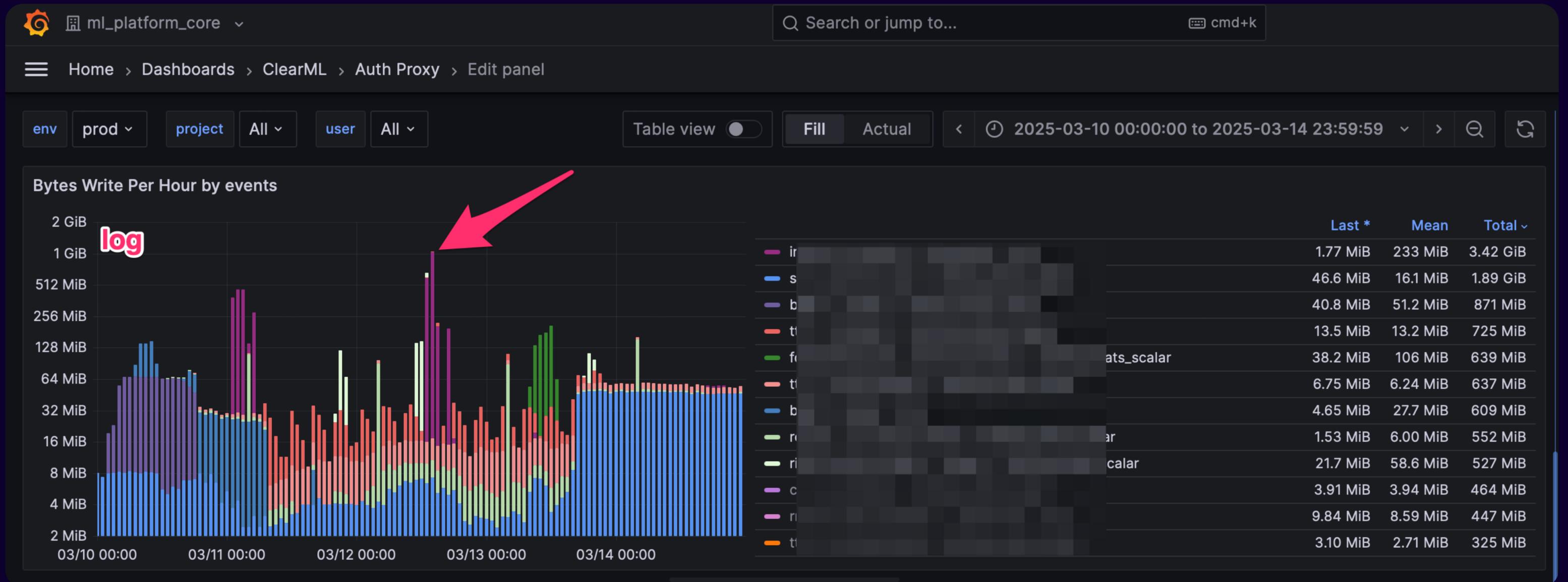
# Тулинг. Трекинг экспериментов / ClearML

Экспериментов в месяц



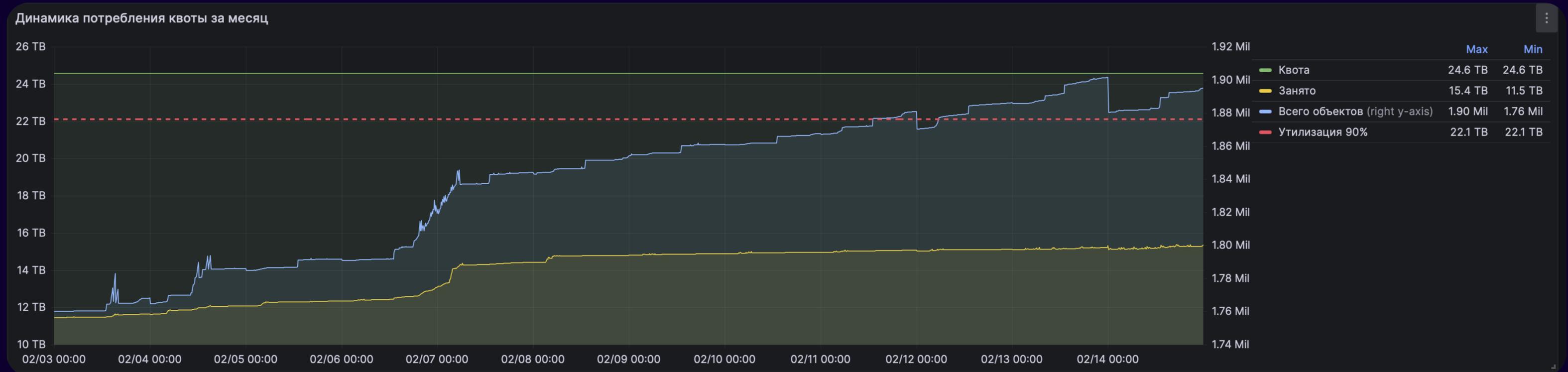
# Тулинг. Трекинг экспериментов / ClearML

Трафик метрик экспериментов по часам



# Тулинг. Трекинг экспериментов / ClearML

Расход файловых хранилищ



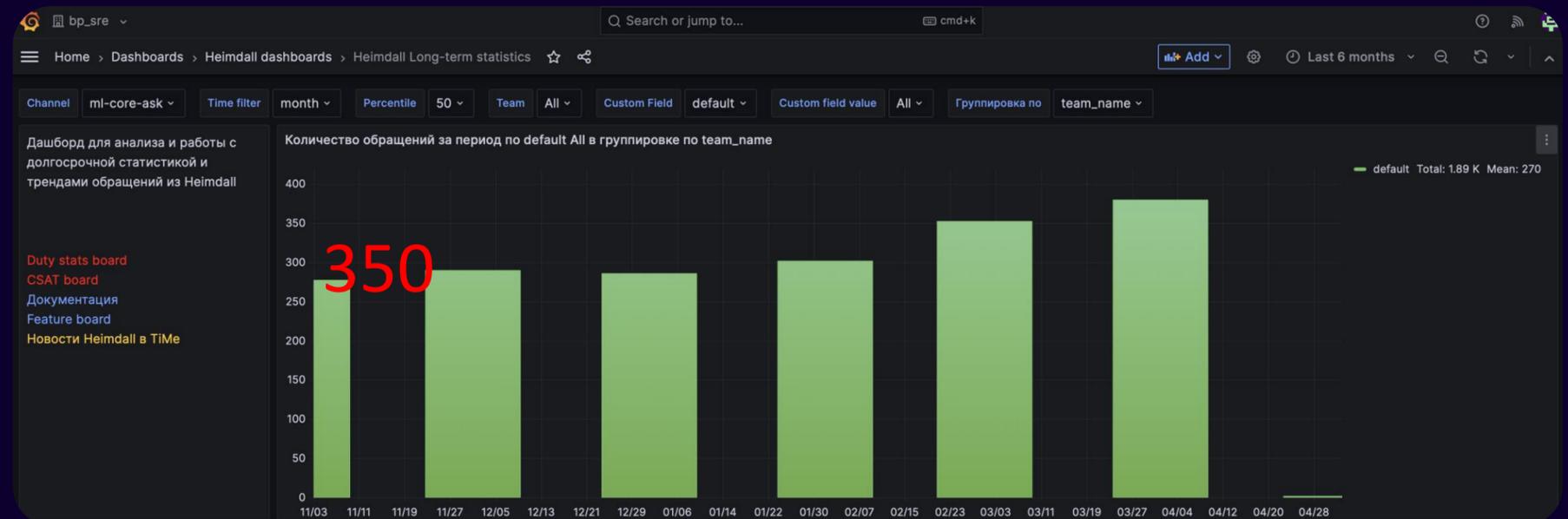
# Тулинг. Трекинг экспериментов / ClearML

- Бесшовная интеграция с Jobs / Jupyter / ModelRegistry / Serving
- Самописная мультитенантность
- +- Кастомизируется
- Держит нагрузку и масштабируется
- Rate limiter и retention (!)
- Тех. независимость / in-house
- Есть приколы... (файлсервер, клиент)



# Инсайты. Поддержка пользователей

- Dev2Dev – это не просто
- 50+% ресурсов команды на RUN-процессы
- Много документации
- 350+ обращений в месяц
- Атмосфера



Обращений в поддержку по месяцам

# А оно надо?

- Кто вы?
  - Сколько вас?
  - Облака тоже круто
  - Много интеграций с другими системами компании
  - Контроль
- 
- Платформа vs SSH-тачки профит: x2

# Вопросы?





Т-Банк — финансовая онлайн-экосистема, объединяющая полный спектр  
финансовых услуг для частных лиц и бизнеса