tech

AI-движок бизнеса:

как единая облачная среда VK Cloud объединяет данные, ML и вывод моделей в прод

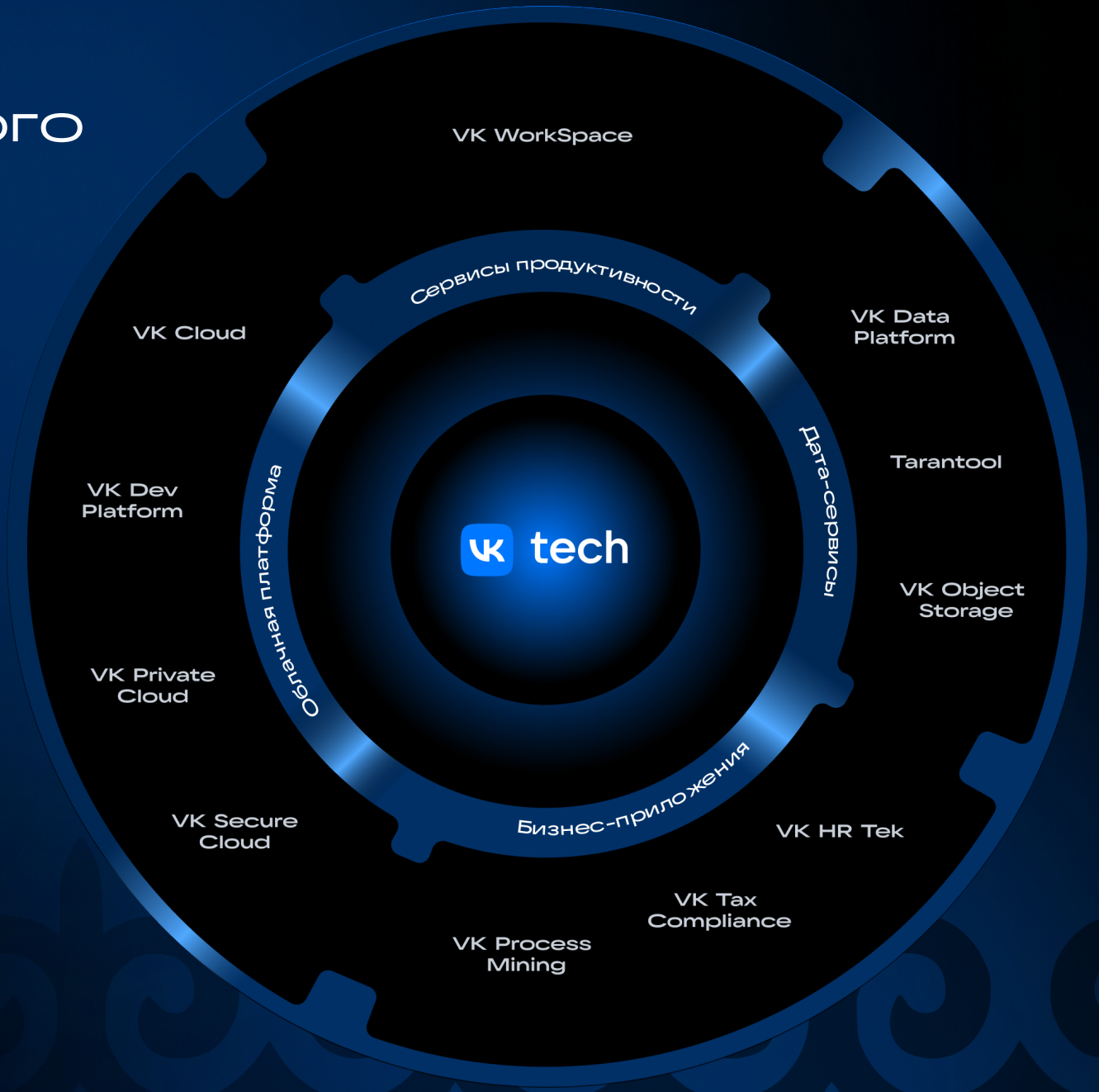
Тлек Адамжанов,
Пресейл архитектор VK Cloud

технологический партнер современного бизнеса

8 лет
с момента создания продукта VK Cloud

>8000 компаний в СНГ
строят бизнес на решениях VK Cloud

300+ инженеров
в разработке облачной платформы



tech

Что мы предлагаем в Казахстане

- 2 зоны доступности
  в г. Астана и г. Косшы
- Все расчеты
  в тенге
- Соответствие закону
  94-V о персональных
  данных
- Сертификат
  PCI DSS
- Юр. лицо — ТОО
  «ВК Тех Казахстан»
- Локальная команда
  в Астане и в Алматы
- Личный кабинет
  на казахском
- Сертификаты ISO 27001,
  27017, 27018, 27701

tech


Астана
ЦОД
АО КАЗТЕЛЕПОРТ (AZ2)
ЦОД
ТОО «QazCloud» (AZ1)
TALAPKER
AKMOLA REGION
ASTANA
ASTANA
AKMOLA REGION
SOUTH-EAST
MICRO-DISTRICT
(RIGHT SIDE)
ОҢТҮСТІК-ШЫҒЫС
ШАҒЫН АУДАН
(ОҢ ЖАҒЫ)
URKER
TOWNSHIP
YPKEP
ШАҒЫН АУДАН
MICRODISTICT
JAGHALAU-3
ЖАҒАЛАУ-3
ШАҒЫН АУДАНЫ
INTERNATSIONALNYI
TOWNSHIP
ИНТЕРНАЦИОНАЛЬНЫЙ
КЕНТІ
QARAOTKEL
TALDYKOL'
ZARECHNOE
ЗАРЕЧНОЕ
TELMAN
TOWNSHIP
ТЕЛЬМАН КЕНТІ
ASTANA
AKMOLA REGION
AKMOLINSKIY
KOSKOPA
AN-MOLA
LESNAYA
POLYANA
ЛЕСНАЯ
ПОЛЯНА

VK Cloud — единая облачнаясреда для данных и AI

Инфраструктурный слой
GPU, IaaS

01

Оркестрация и абстракция ресурсов
Kubernetes, Autoscaling,
Namespace / quota / multi-tenancy

02

Слой данных (Data Layer)
Object Storage (data lake),
DWH / аналитические базы,
Стриминг (Kafka)

03

Data Processing & Analytics
ETL / ELT, Spark / Flink

04

ML / AI Platform (Model Development)
JupyterHub, MLFlow

05

MLOps / Model Serving (Model Development)
MLDeploy

06

AI Applications / Business Layer
AI-сервисы, LLM-приложения,
Чат-боты

07

Cross-cutting слои (поперечные, сквозные)
Security / IAM, Governance
& compliance, Monitoring & logging,
Cost management

08

tech

Высокая эффективность в области ИИ на базе HGX Платформы с NVIDIA H200 SXM

Это инвестиции в **скорость, эффективность и масштабируемость**.
Получите стратегическое преимущество которое гарантирует, что компания будет не просто идти в ногу со временем, а определять темп развития.

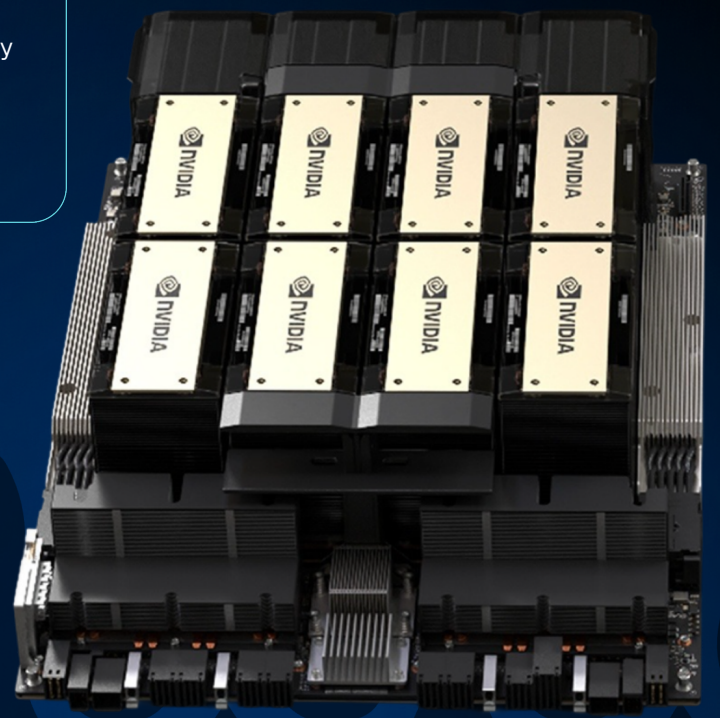
Виртуализация с GPU

- В обучении генеративных ИИ-моделей, инференсе мультимодальных моделей наблюдается прирост **производительности до 43%** в сравнении с H100
- Возможность запустить мощные open-source **LLM DeepSeek-671B и Qwen3-235B** и ускорять Time to market на существующих Llama, GPT, Gemma, Mixrtal и др.
- Решает ключевую проблему современного ИИ — ограничение по размеру модели и пакетов данных за счёт **максимально доступного совокупного объема GPU-памяти**
- Оснащена полностью связанной фабрикой NVSwitch, обеспечивающей пропускную способность 7.2 ТБ/с «все-ко-всем» между GPU, что **в 8 раз выше**, чем у конфигурации с мостами NVLink на базе PCIe3
- H200 – **первая GPU с памятью HBM3e**, обеспечивающая 141 ГБ емкости (в 1,76 раза больше, чем у H100) и 4.8 ТБ/с пропускной способности (увеличение в 1.4 раза по сравнению с H100)
- HGX обеспечивает связанность 8 GPU-карт без узких мест, **обеспечивая дополнительную производительность** в сравнении с H200 на PCIe

Cloud Containers (k8s) с GPU

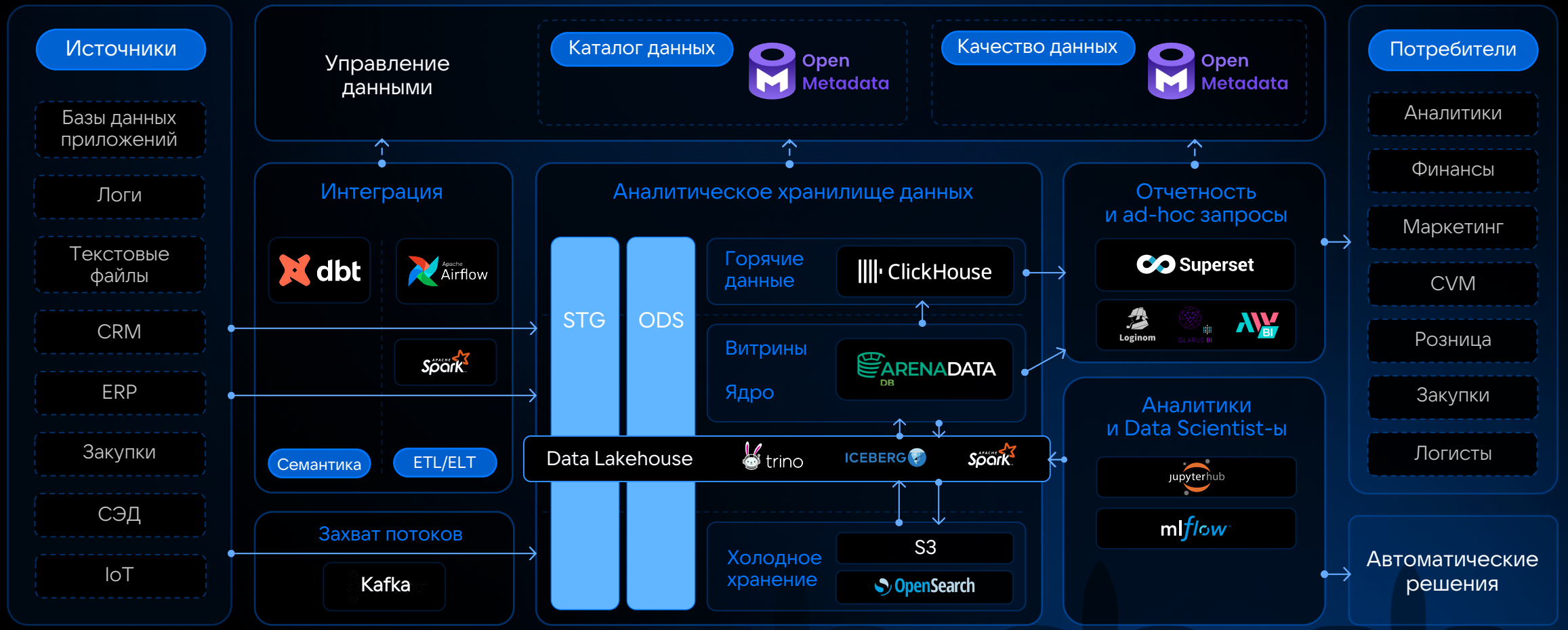
- Под любую вашу задачу у нас есть подходящая GPU в Cloud Containers
- Не нужно тратить ресурсы на ручное масшабирование, наш сервис Cloud Containers автоматически создаст ноды и распределит нагрузку
- Одна GPU может обслуживать разные команды data scientist в изолированных инстансах, выполняя множество задач одновременно и экономя время

Конфигурация:
8- GPU h200 SXM5 с NVLink + Quantum-2 InfiniBand.
CPU 2× AMD EPYC™ 9654 (192 ядра / 384 потока)
RAM: 2.3+ ТБ DDR5 (24× 96 GB), частота 5600 Mhz



tech

Слой данных (Data Layer) и Data Processing



tech

ML / AI Platform (Model Development) и MLOps

Рабочее место ML-инженера

Среда для обучения моделей, экспериментов и написания кода

Версионирование моделей

Позволяет отслеживать метрики качества и параметры обучения

Деплой моделей

Превращает обученную модель в готовый сервис

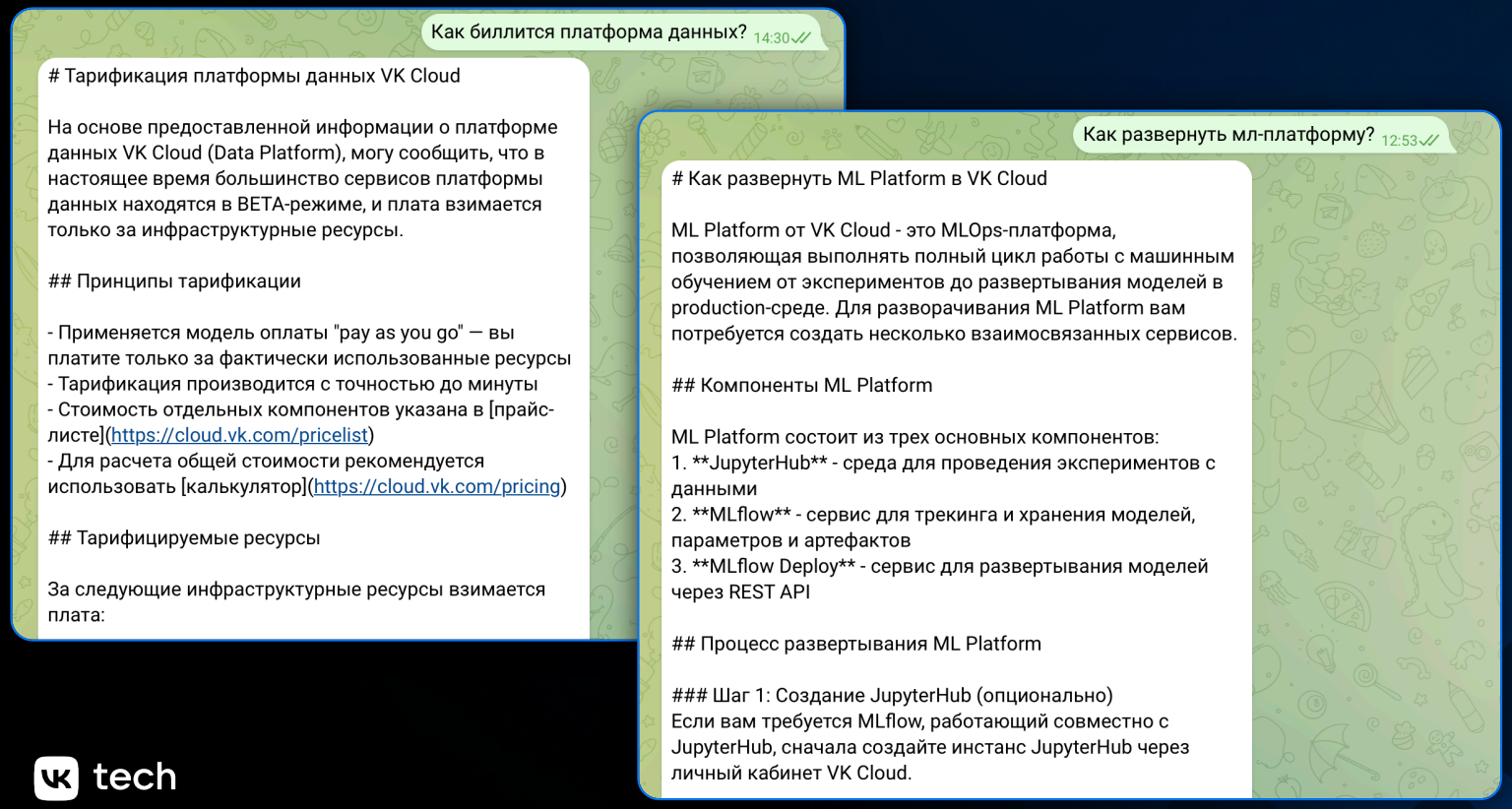
Большие данные

Для изучения данных, расчета фич, формирования витрин

Кейс: бот с доступом к базе документов

Пользователь пишет вопрос на естественном языке

Бот ищет релевантные документы и формирует ответ



ИИ-консультант VK Cloud

tech

Кейс: почтовый робот для helpdesk в FMCG-компании

Пользователь отправляет письмо на почту IT Support

Агент анализирует запрос и формирует ответ на основании релевантных документов

Бот формирует подробный ответ на естественном языке.

При необходимости бот может перенаправить запрос специалисту:

- Если бот не уверен в своем ответе или если нет данных
- Если требуются активные действия (сброс пароля, создание заявки и т. п.)

Инструкции для пользователей

Историческая база писем

tech

AI-консалтинг

Адаптируем лучшие AI-модели под задачи ваши бизнеса.
Полное сопровождение: от идеи и постановки задачи до внедрения и обучения сотрудников.

- Подбираем и адаптируем лучшие open-source модели
- Разрабатываем готовый сервис
- Фокус на бизнес-результате

1 этап
Анализ бизнес-задачи, погружение в производственные процессы

2 этап
Разработка архитектуры решения, наилучшим образом соответствующего потребностям бизнеса

3 этап
Разработка и внедрение AI-решения

4 этап
Сбор обратной связи и анализ бизнес-метрик

5 этап
Техническая поддержка и обучение сотрудников


Погружение в процессы, сбор данных и обоснование

Разработка

Внедрение и сопровождение

tech

Обучение от VK Cloud

Построение и эксплуатация AI/ML-решений в Enterprise на базе VK Cloud

0₽ Бесплатный воркшоп от VK Cloud:

Как спроектировать, развернуть и масштабировать AI-решения — от PoC до production

tech

tech

Рахмет!

Тлек Адамжанов,
Пресейл архитектор VK Cloud

t.adamzhanov@vkteam.ru