



Приоритет 2030

Выездной семинар
**“Модернизация образовательных программ по
программной инженерии в условиях научно-
технологического развития России”**

SmartMLOps - Корпоративная информационная система
поддержки моделей искусственного интеллекта НИУ ВШЭ



Приоритет 2030

ИИ-технологии для человека

SmartMLOps - Корпоративная информационная система поддержки моделей искусственного интеллекта НИУ ВШЭ

Салех Хади Мухаммед, Департамент программной инженерии
Факультета компьютерных наук НИУ ВШЭ г. Москва
hsalekh@hse.ru

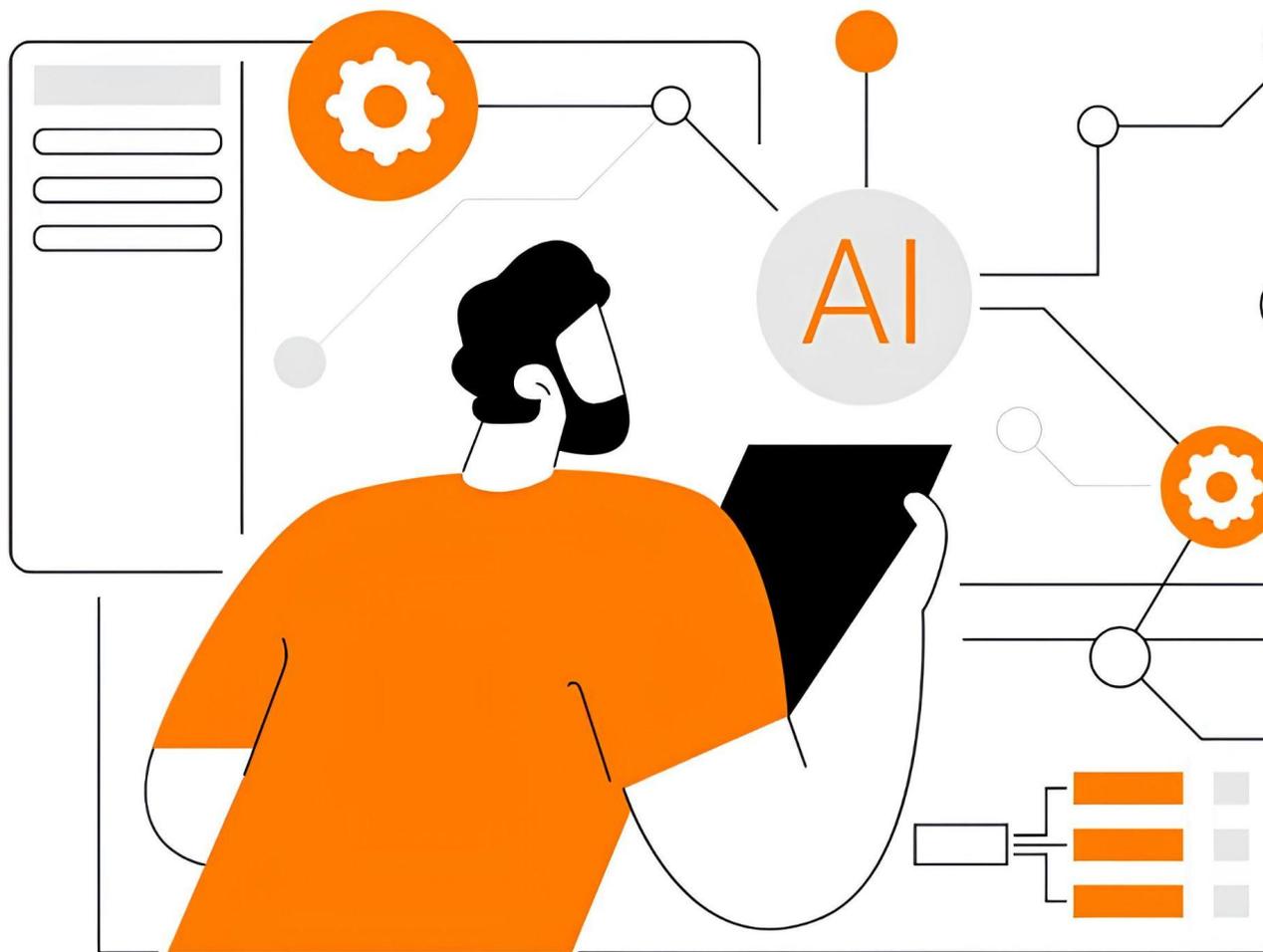
Москва 2025



Целевая аудитория

MLOPS платформа нацелена на разработчиков ИИ и ML моделей желающих масштабировать свои интеллектуальные сервисы без необходимости решать инфраструктурные задачи и задачи безопасности.

- Лаборатории ВШЭ
- Проекты СП4, СП5
- Подразделения ВШЭ
- Студенты, аспиранты, ППС и сотрудники ВШЭ
- Внешние исследователи и партнеры



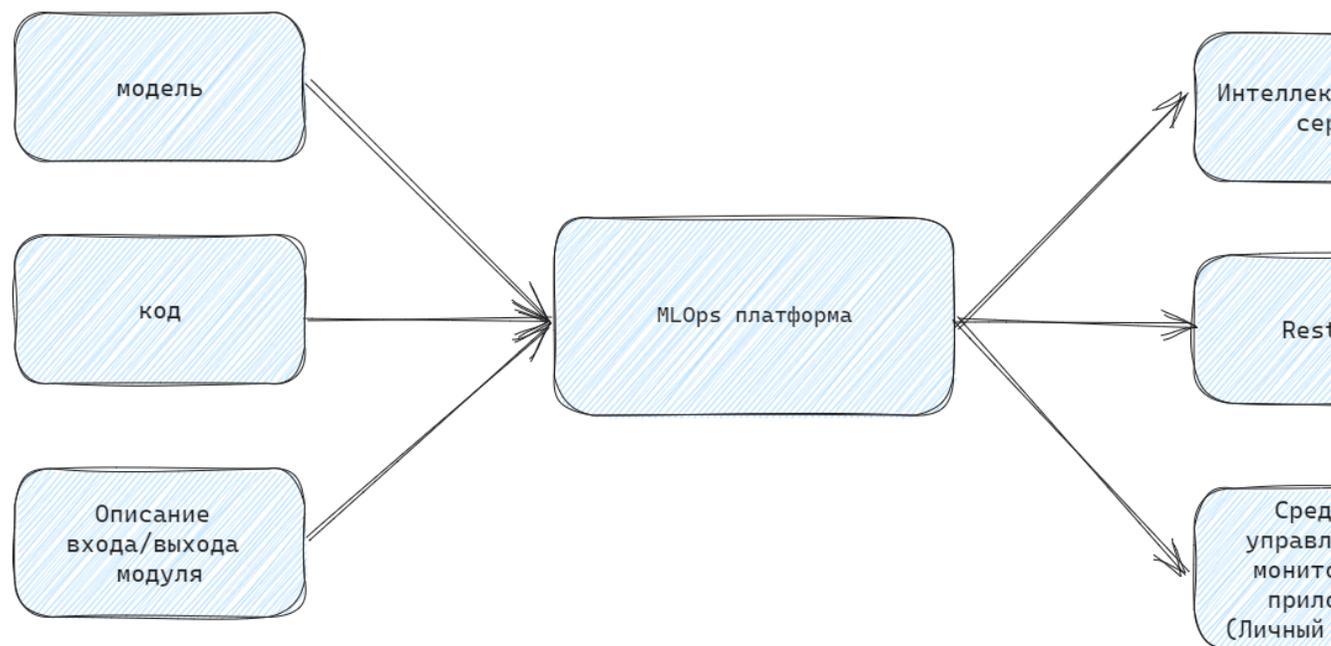


MLOps платформа

Что такое MLOps платформа:

- вычислительные ресурсы
- хранилища данных
- система безопасности
- платформа для пользователей
- Доступ к суперкомпьютеру CHARISMa
- Мониторинг ресурсов и работы моделей
- Управление доступом и лицензиями к модулям
- Консультации и сопровождение от команды техподдержки

для вашего интеллектуального приложения





Что уже сделано 2024



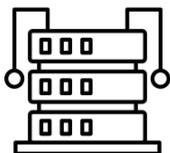
Анализ конкурентных
решений



Развертывание фреймворка
в облаке и настройка
инфраструктуры



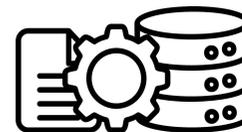
Развертывание ИИ
моделей на платфор



Интеграция
платформы с
суперкомпьютером



Научные
семинары



Интеграция КИС НИУ
ВШЭ + ETL



Реализация
аналитических пан



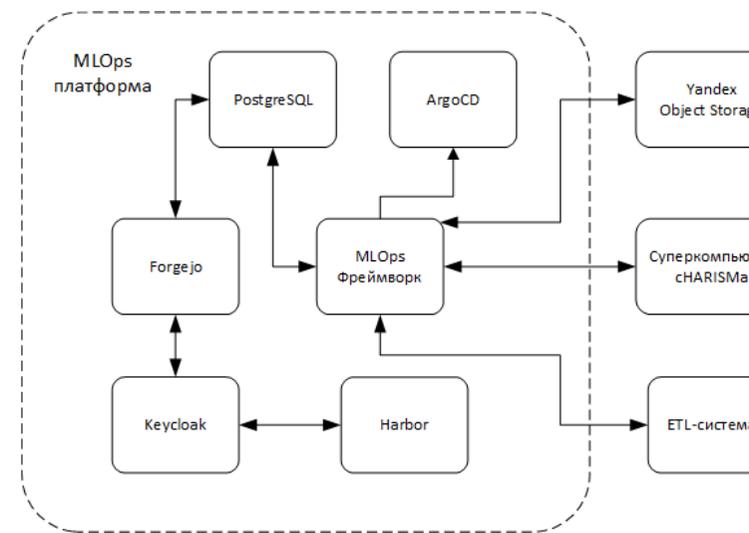
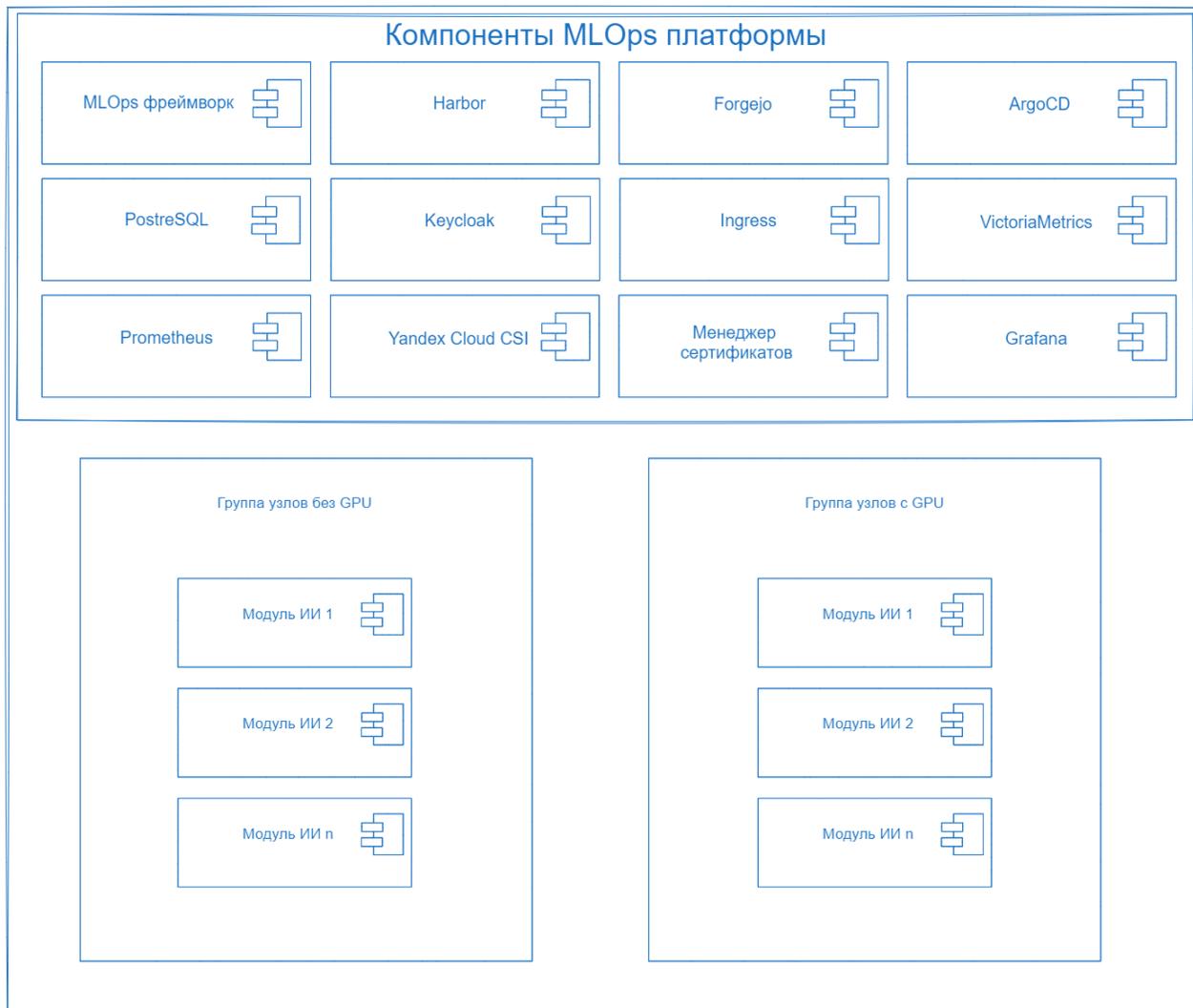
РИД



Безопасность

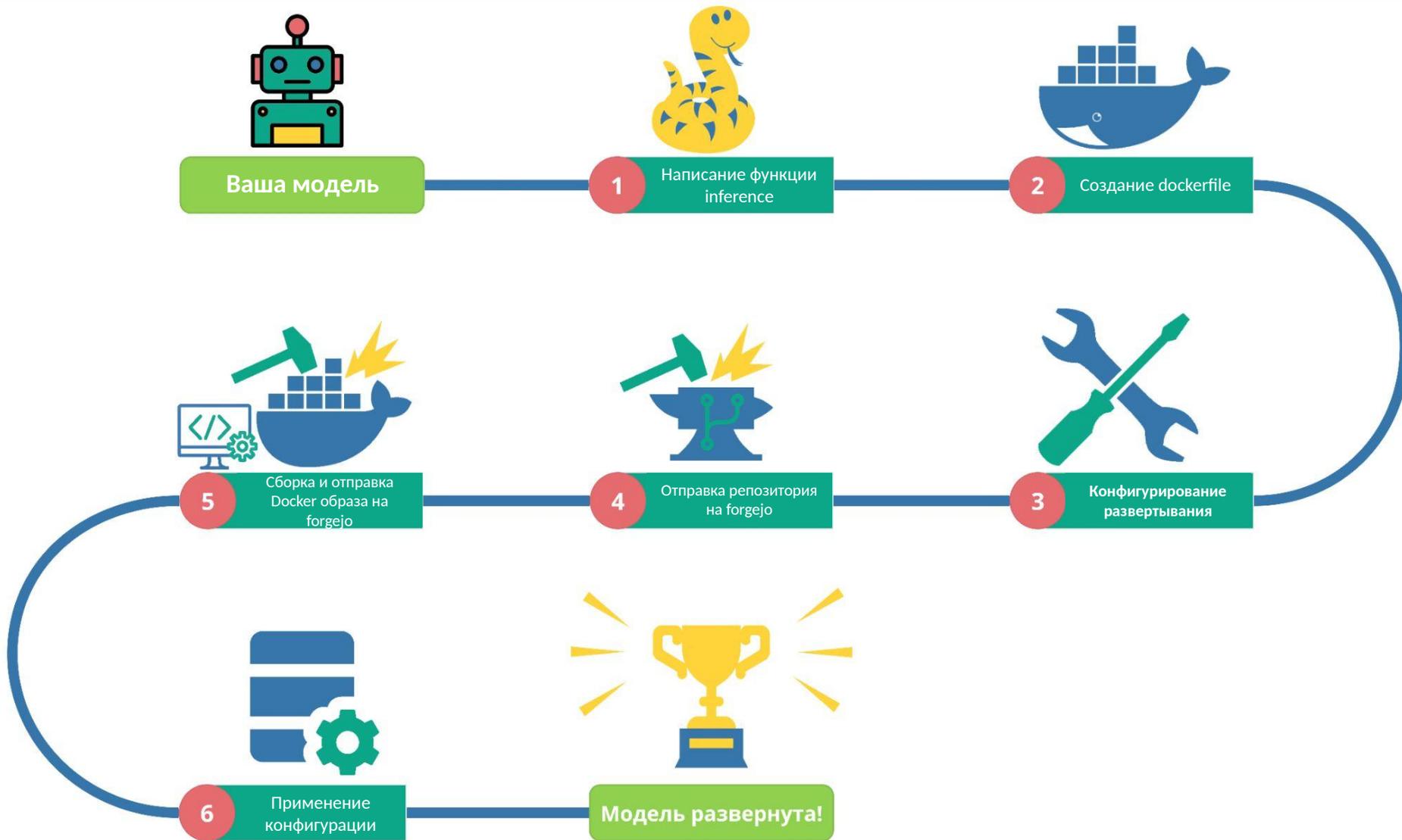


Инфраструктура платформы



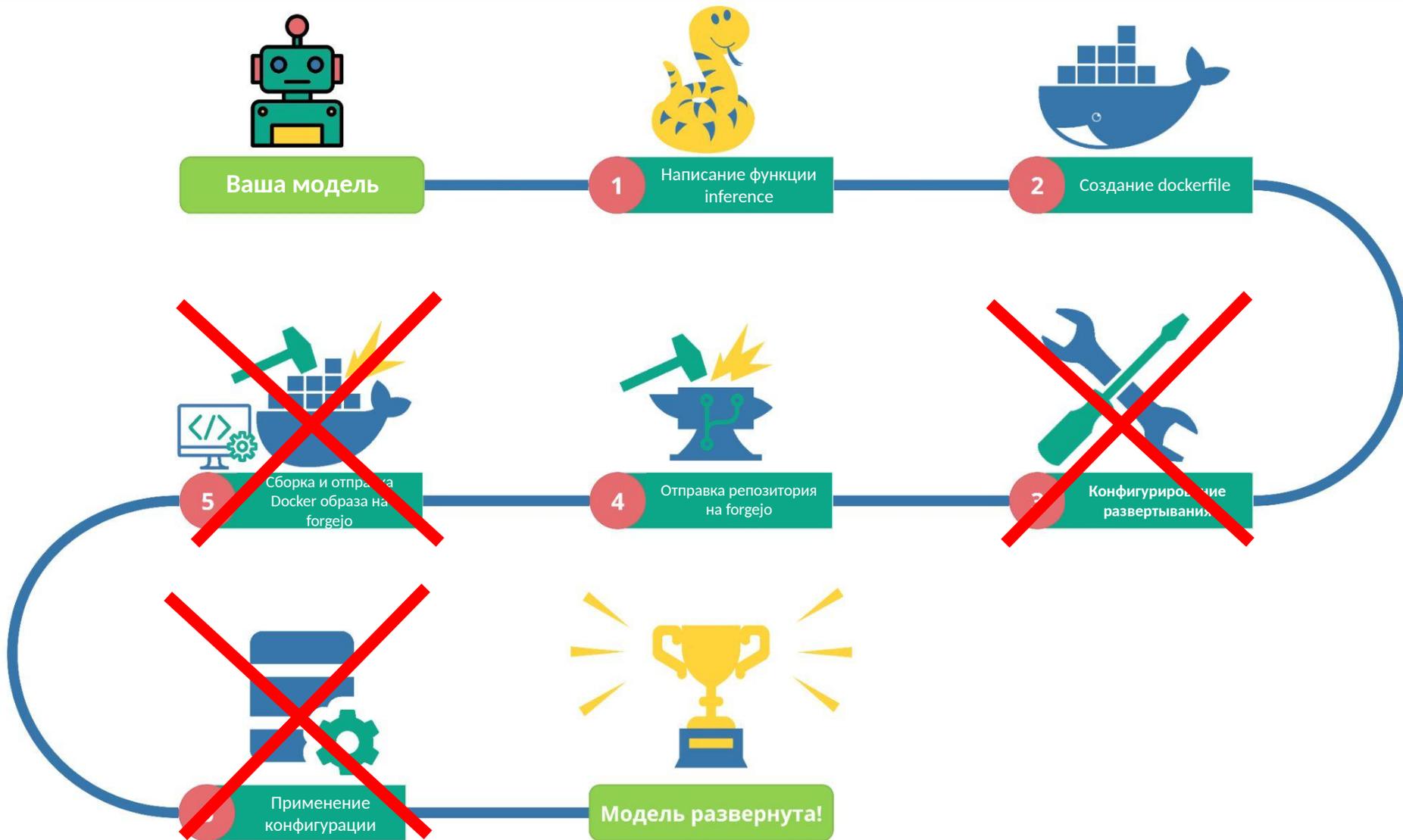


Развертывание моделей на MLOPS платформе в 2024





Развертывание моделей на MLOPS платформе в 2025

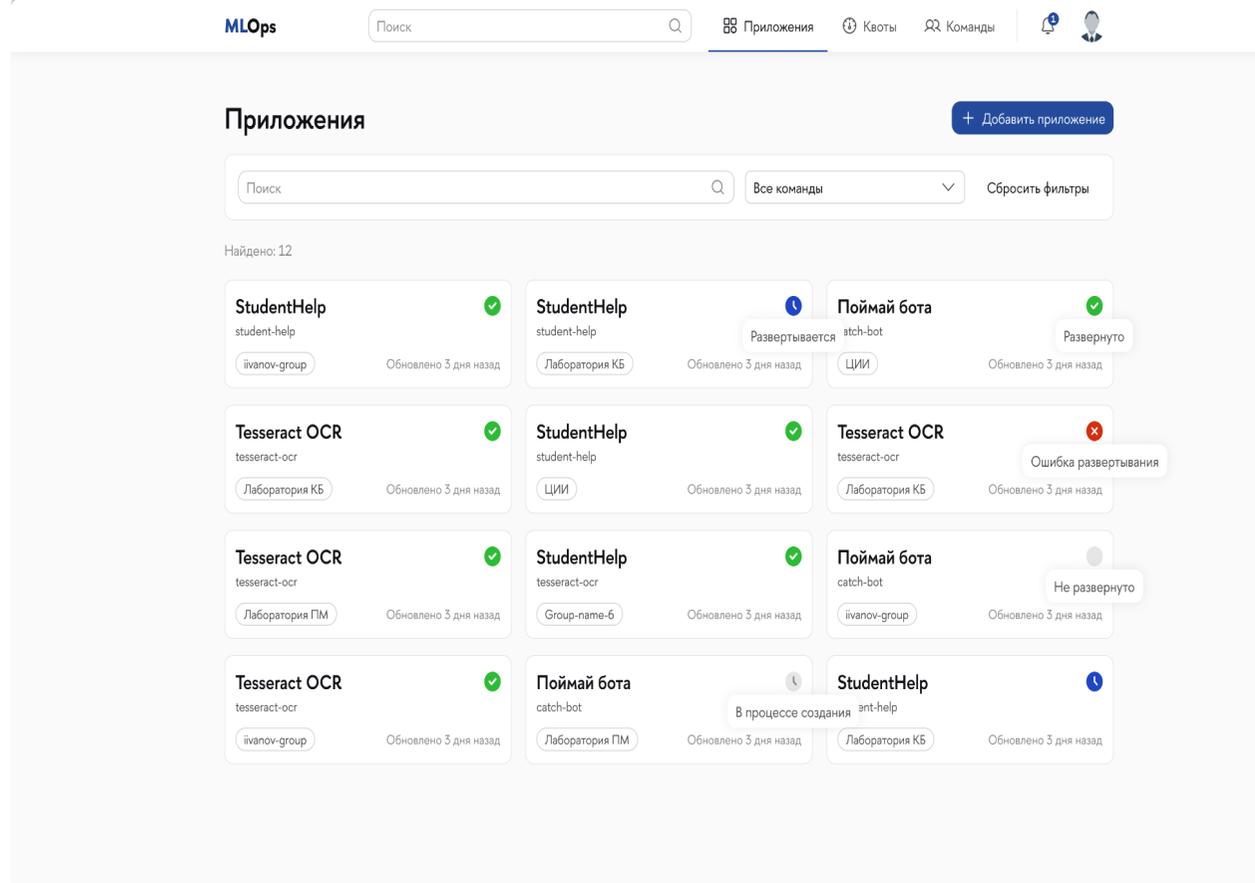


Личный кабинет разработчика

Функционал ЛК:

- Упрощение процесса создания и управления приложением:
 - Генерация и шаблонное заполнение манифестов
 - Логи развертывания
 - Управление компонентами
 - Статусы
- Мониторинг
 - Дашборды разработчика
- Разграничение доступа
 - OIDC

Работа над Личным кабинетом не окончена! Внешний вид интерфейса может быть изменен.

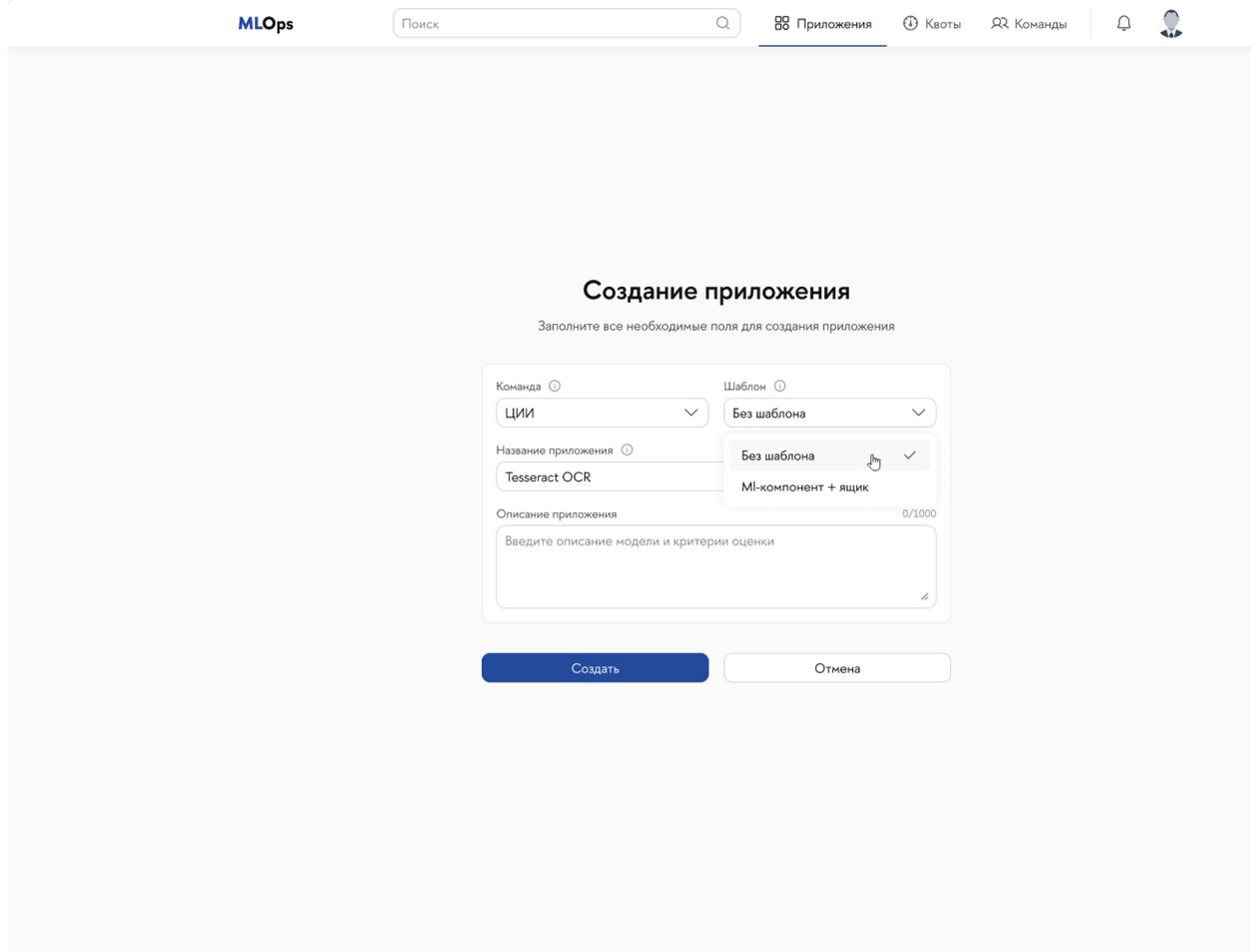


Создание приложения через Личный кабинет

Меню создания приложения:

- Имя приложения
- Выбор команды разработчиков поддерживающих приложение
- Выбор заранее созданного командой шаблона приложения
- Краткое описания для отображения в каталоге

Работа над Личным кабинетом не окончена!
Внешний вид интерфейса может быть изменен.



The screenshot shows the 'Создание приложения' (Create Application) form in the SmartMLOps system. The form is titled 'Создание приложения' and includes the instruction 'Заполните все необходимые поля для создания приложения'. The form fields are:

- Команда (Team):** A dropdown menu with 'ЦИИ' selected.
- Шаблон (Template):** A dropdown menu with 'Без шаблона' selected. A tooltip is visible over this dropdown, showing two options: 'Без шаблона' (selected) and 'ML-компонент + ящик'.
- Название приложения (Application Name):** A text input field containing 'Tesseract OCR'.
- Описание приложения (Application Description):** A text area with a character count of '0/1000' and the placeholder text 'Введите описание модели и критерии оценки'.

At the bottom of the form, there are two buttons: 'Создать' (Create) and 'Отмена' (Cancel).

Создание приложения через Личный кабинет

После создания разработчик получает:

- Организацию в Harbor
- Репозиторий приложения в Fogejo
- Мониторинг Grafana

Список компонент приложения

Контроль доступа к приложению

Возможность добавить новые компоненты

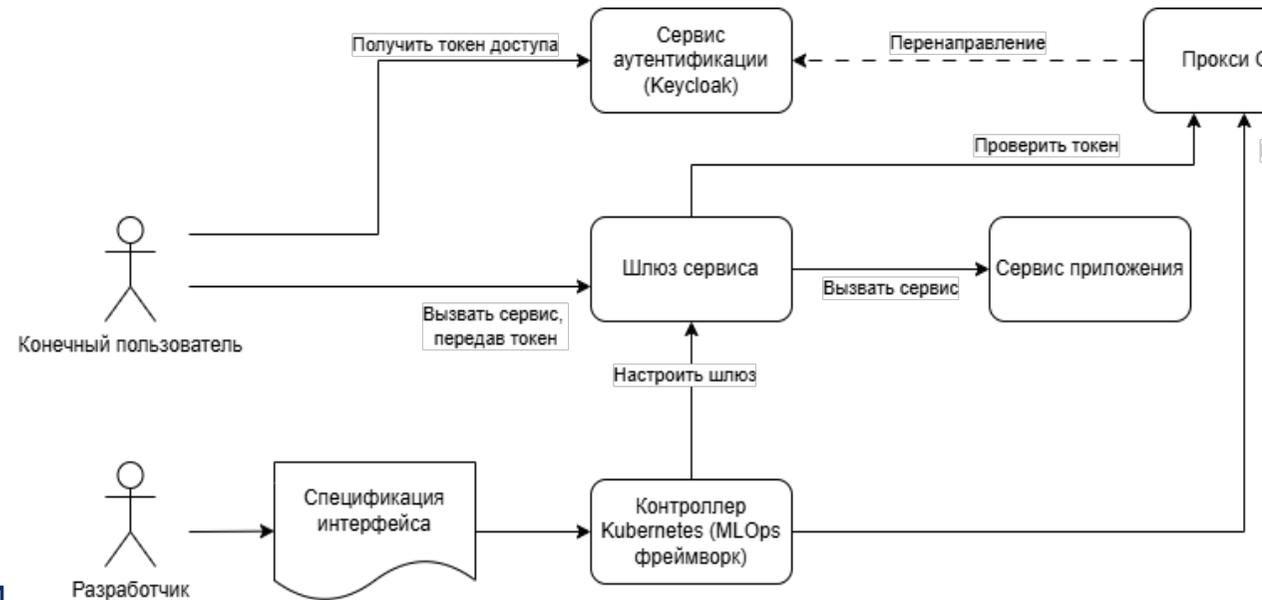
Работа над Личным кабинетом не окончена!
Внешний вид интерфейса может быть изменен.

The screenshot displays the MLOps application interface for configuring a 'Tesseract OCR (tesseract-ocr)' application. The top navigation bar includes the MLOps logo, a search bar, and links for 'Приложения', 'Квоты', and 'Команды'. The main content area shows the application name with a green checkmark, update information ('Обновлен 8 июля 2024, 12:13'), and the author ('Автор Иван Крупнов'). Below this are three tabs: 'Компоненты', 'Настройки', and 'Доступы и безопасность'. The 'Компоненты' tab is active, showing a list of component types: 'ML-компоненты', 'Пайплайны' (7), 'Ящики', and 'API-компоненты'. A '+ Добавить ML-компонент' button is visible in the top right of the component list area. The main content area below the tabs is currently empty, displaying 'Пока пусто' and a message: 'Когда будут добавлены ML-компоненты, они будут отображаться здесь'.

Появилась аутентификация на основе OIDC

Функция аутентификации и защиты доступа платформы предоставляет возможности:

- для разработчиков единая точка входа (Single Sign-On) во всех сервисах платформы;
- централизованное управление доступом разработчиков и конечных пользователей к сервисам платформы и сервисам размещенных на платформе приложений;
- разработчики сами могут управлять доступом к размещенным им приложениям;
- подключение внешних служб идентичности и каталогов пользователей;



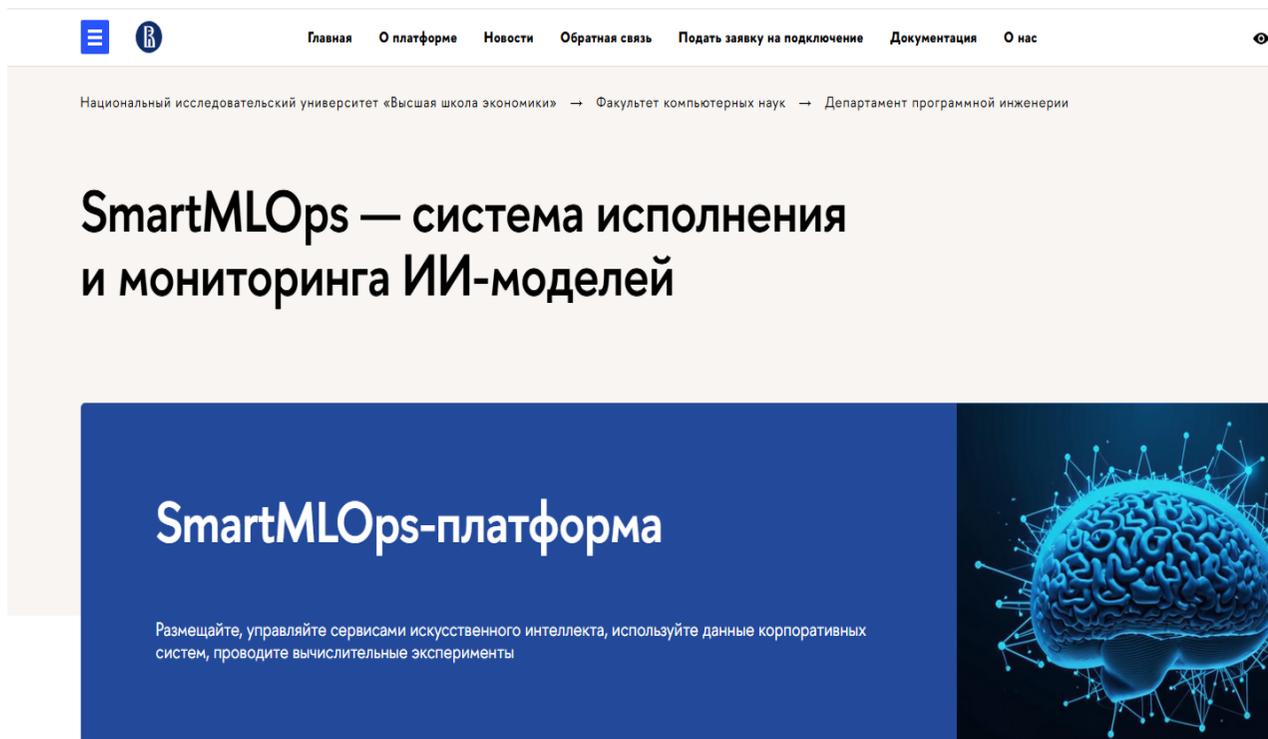
Появился Вэб портал MLOPS

Точка входа пользователя в систему MLOps:

- Подача заявки для подключения
- Новости платформы
- Документация и гайды
- Обратная связь
- Тех. поддержка



<https://mlops.hse.ru/>



The screenshot shows the web portal interface for SmartMLOps. At the top, there is a navigation menu with links: Главная, О платформе, Новости, Обратная связь, Подать заявку на подключение, Документация, О нас. Below the menu, the breadcrumb path is: Национальный исследовательский университет «Высшая школа экономики» → Факультет компьютерных наук → Департамент программной инженерии. The main heading reads: SmartMLOps — система исполнения и мониторинга ИИ-моделей. Below this, there is a blue banner with the text: SmartMLOps-платформа. Underneath the banner, it says: Размещайте, управляйте сервисами искусственного интеллекта, используйте данные корпоративных систем, проводите вычислительные эксперименты. On the right side of the banner, there is an image of a brain with neural connections.

Важное объявление

Приглашаем Вас принять участие в опытной эксплуатации в период с 01.02.2025 по 31.12.2025!



Документация

Пользовательская документация:

- Размещена в общедоступном репозитории на forgejo
- Все еще в работе и постоянно дополняется
- Контент доступный на данный момент(На изображении справа):
- В том же репозитории можно сообщить о баге, предложить идею, задать вопрос о деталях работы платформы:

MLOps Платформа

- Базовые темы
 - Введение во фреймворк
 - Работа с фреймворком, основные компоненты
 - Размещение приложения на основе шаблона
- Подробно
 - API модулей
 - Компонент DataBox и работа с файлами
 - Компонент MLComponent
 - Компонент ExperimentPipeline
 - Терминология

Репозиторий доступен на портале или по ссылке



https://platform-forgejo.stratop.com/mlops_platform/documentation

Метки Этапы Поиск задач... Приблизительный 🔍 Создать задачу

🕒 4 Открыто ✓ 0 Закрото Метка Этап Проект Автор Назначено Тип Сортировка

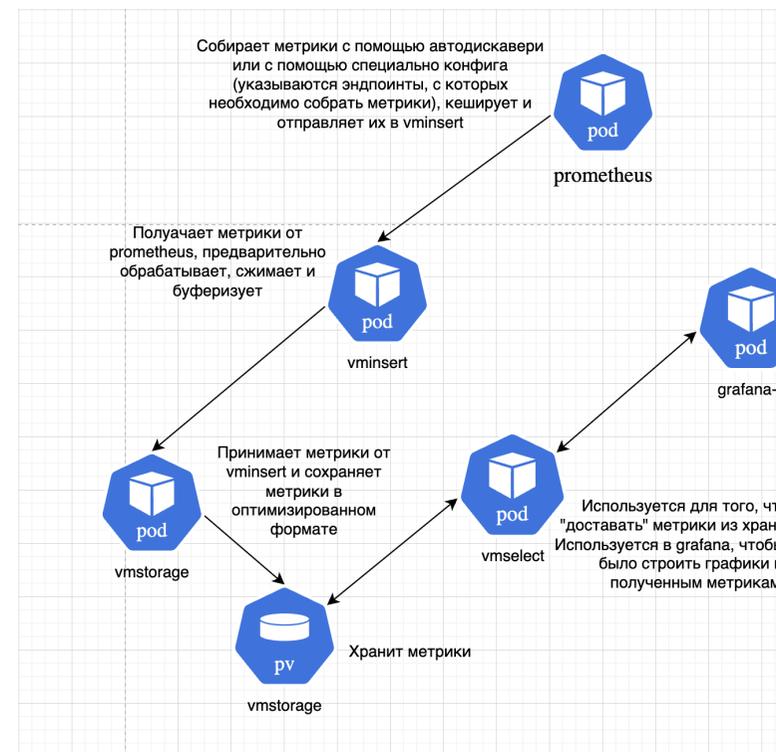
- 🟢 В хранилище S3 создаются пустые папки, если пайплайн ссылается на несуществующий файл **bug**
#4 открыта 8 часов назад gzhulikov
- 🟢 При удалении приложения некоторые компоненты не удаляются **bug**
#3 открыта 4 дня назад gzhulikov
- 🟢 API компоненты пайплайнов не работают сразу после создания **bug**
#2 открыта на прошлой неделе gzhulikov
- 🟢 Удаление развернутого приложения
#1 открыта на прошлой неделе maksgalinchik

Система мониторинга платформы: сделано в 2024

Собраны базовые дашборды для администратора системы

Состав дашбордов:

- Состояние пространств имен Kubernetes
 - Просмотр состояния по нагрузке CPU и использованию памяти
- Состояние по нагрузке на Kube API server
- Мониторинг состояния Ingress контроллеров
- Мониторинг состояния СУБД
- Состояние системы бэкапов Velero
- Состояние системы хранения VictoriaMetrics
- Состояние выполнения задач на кластере CHARISMa



Система мониторинга платформы: план на 2025

Разработка персональных дашбордов для разработчиков

Цель: упрощение процесса мониторинга состояний приложений для разработчиков

Планируемый состав:

- Мониторинг ресурсов (vCPU, Ram, SSD) пространств имен приложение в Kubernetes
- Мониторинг журналов приложения
- Мониторинг журнала развертывания приложения
- Мониторинг сетевой нагрузки приложения
- Мониторинг журналов из Ingress связанные с приложением
- Формирование отчета синхронизации Argo CD для приложения
- Мониторинг состояния Harbor
- Мониторинг состояния Forgejo
- Мониторинг состояния S3

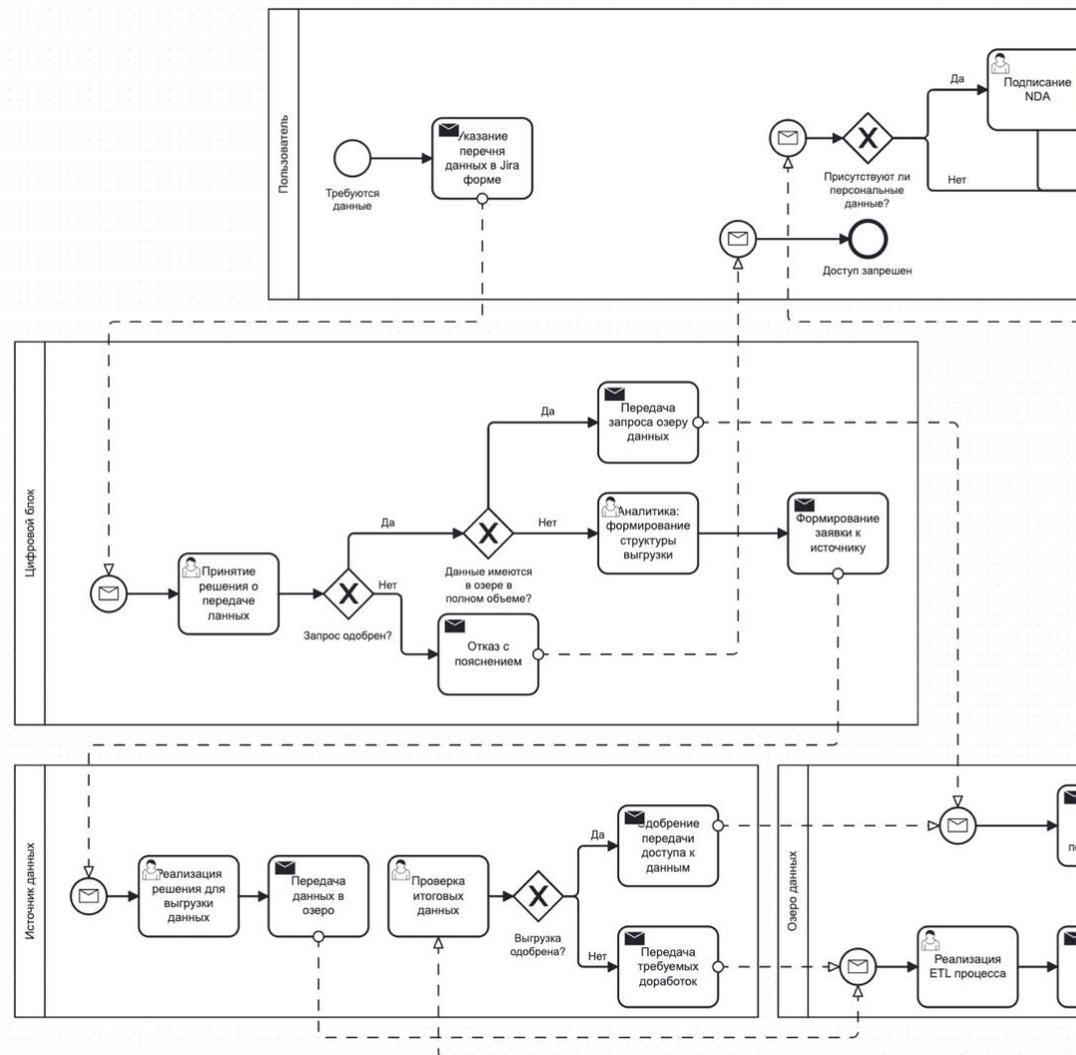
Расширение состава дашбордов для администратора системы

- Мониторинг состояния журналов событий всех пространств имен всех пользователей
- Мониторинг журнала развертывания приложений
- Мониторинг состояния Harbor
- Мониторинг состояния Forgejo
- Мониторинг состояния S3
- Мониторинг текущих выставленных лимитов в приложении



Озеро данных

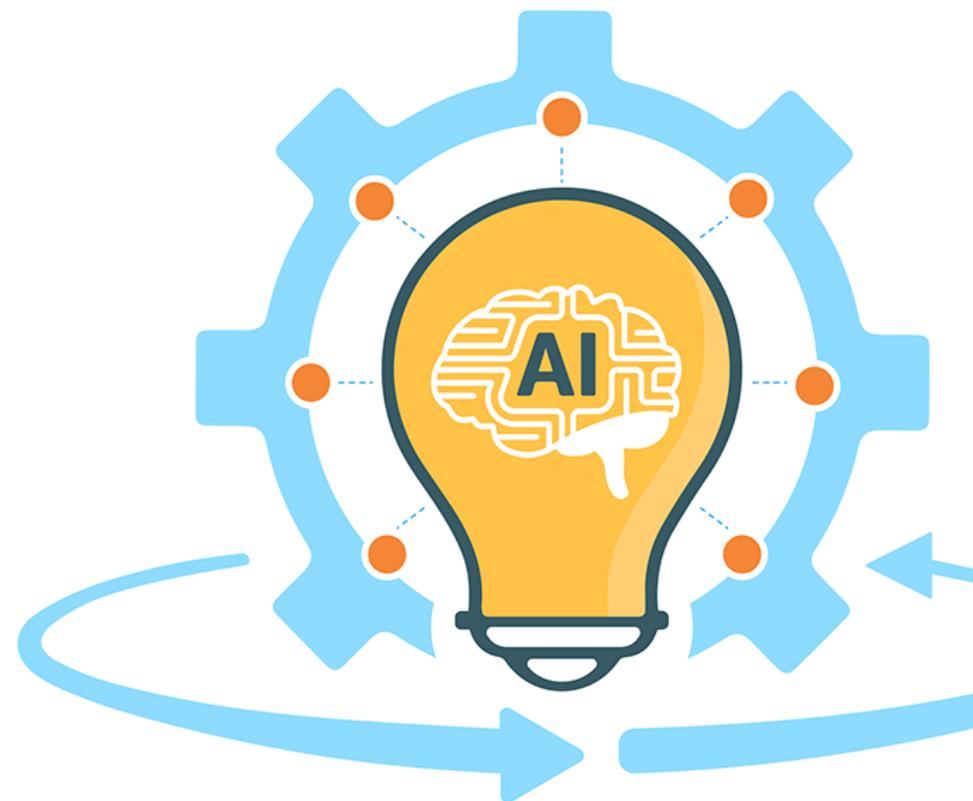
- Необходимость централизованного хранения данных:
 - повторная реализация механизма выгрузки
 - Дублирование данных и неоднократные обращения к источникам
 - Сложности с единым управлением доступом к различным наборам данных
 - Необходимость обезличивания данных
- Появление Озера Данных даст:
 - Возможность переиспользования
 - Данные заранее очищены и структурированы
 - Сокращение трудоемкости процессов интеграции





План на 2025

- Личный кабинет разработчика ИИ моделей
- Конструктор пайплайнов в ЛК разработчика
- Мониторинг моделей для разработчика
- Мониторинг платформы для администраторов
- Биллинг
- Информационная безопасность
- Аутентификация пользователей
- Инструменты обезличивания данных КИС
- Опытная эксплуатация





Прогресс выполнения плана 2025

Личный кабинет разработчика ИИ моделей (Демо)

Конструктор пайплайнов в ЛК разработчика

Мониторинг моделей для разработчика

Мониторинг платформы для администраторов

Биллинг

Информационная безопасность

Аутентификация пользователей (Демо)

Инструменты обезличивания данных КИС

Опытная эксплуатация (Начало)

Спасибо за внимание!

Обратная связь:



Присоединяйтесь к опытной эксплуатации

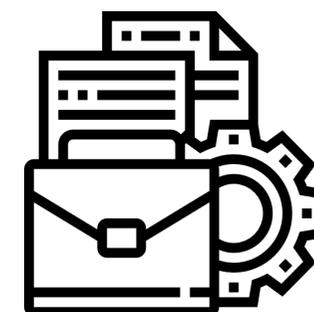


Салех Хади Мухаммед, Департамент программной инженерии
Факультета компьютерных наук НИУ ВШЭ г. Москва
hsalekh@hse.ru



Направление работы 2025

- Улучшение пользовательского опыта:
 - Упрощение процесса создание приложения
 - Личный кабинет
 - Мониторинг
- Развитие средств администрирования
 - Мониторинг администратора
 - Аутентификация пользователей
 - Биллинг
- Развитие платформы
 - Информационная безопасность
 - Инструменты обезличивания данных КИС
 - Конструктор пайплайнов в ЛК разработчика





0

Создание приложения в ЛК

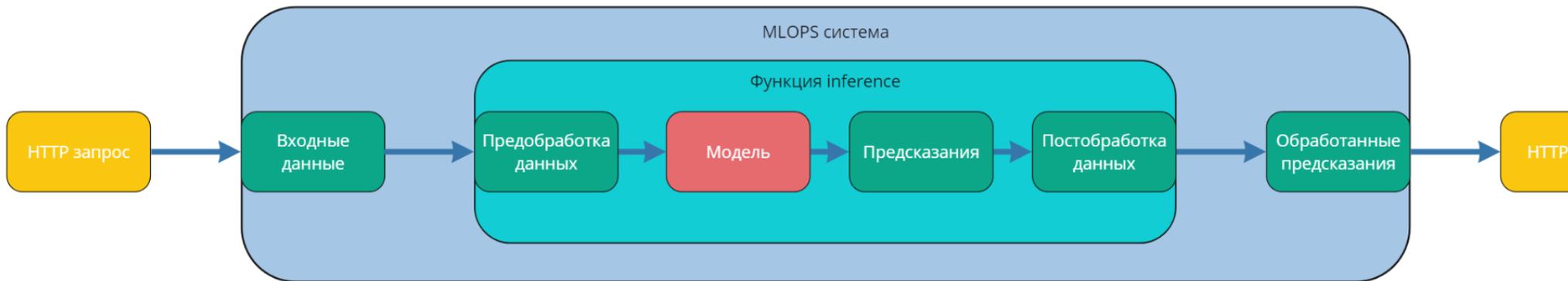
- Каждый компонент обладает экраном событий сигнализирующем о состоянии компонента.
- При переходе в статус можно будет увидеть логи указывающие на причину статуса

Работа над Личным кабинетом не окончена!
Внешний вид интерфейса может быть изменен.

The screenshot displays the MLOps dashboard for a component named 'Tesseract OCR (tesseract-ocr)'. The interface includes a search bar, navigation tabs for 'Приложения', 'Квоты', and 'Команды', and a user profile icon. The component page shows it was updated on 8 июля 2024, 12:13 by author Иван Крупнов. A sidebar on the left lists 'ML-компоненты' (7), 'Пайплайны' (7), 'Ящики', and 'API-компоненты'. The main content area shows the component 'MLComponent_1' with a 'События' (Events) tab selected. The event log contains four entries:

Статус	Время	Действие
Развертывается	3 июня 2024, 19:33	
Ошибка развертывания	3 июня 2024, 19:33	
Развернуто	3 июня 2024, 19:33	
Развертывается	3 июня 2024, 19:33	

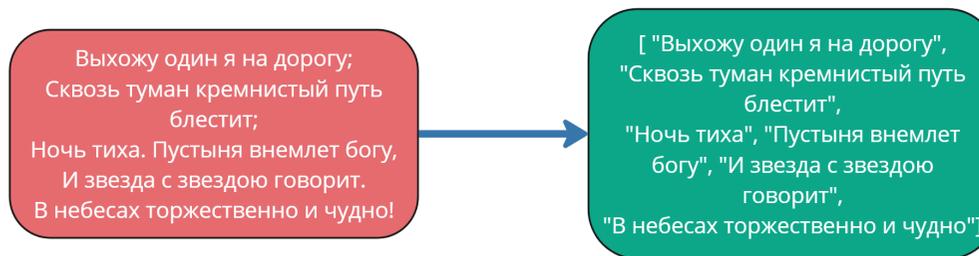
1 Написание функции inference



Функция inference нужна для:

- Чтения входных данных из файла
- Предобработки входных данных
- Передачи входных данных в модель
- Получения результатов работы модели
- Обработки результатов работы модели, приведение данных в нужный формат
- Запись результатов в файл

Пример предобработки





2

Создание dockerfile

- Можно взять шаблон и отредактировать
- Содержимое файла как правило не меняется от модели к модели
- На основе этого файла потом соберется docker образ
- В образ копируются все файлы, необходимые для работы модели (например файл, содержащий функцию inference)
- Сами файлы модели в образ не кладутся, они помещаются в S3 хранилище, но об этом позже

Пример dockerfile

```
FROM python:3.9-slim-bullseye

USER root
RUN useradd -ms /bin/bash app
WORKDIR /home/app

RUN pip3 install --upgrade pip wheel setuptools
COPY requirements.txt ./requirements.txt
RUN pip3 install -r ./requirements.txt

COPY --chown=app . ./
USER app

ENV PATH=$PATH:/home/app/.local/bin
ENTRYPOINT ["python3.9", "-m"]
```

3 Сборка и отправка Docker образа на forgejo

Сборка образа проходит в одну команду:

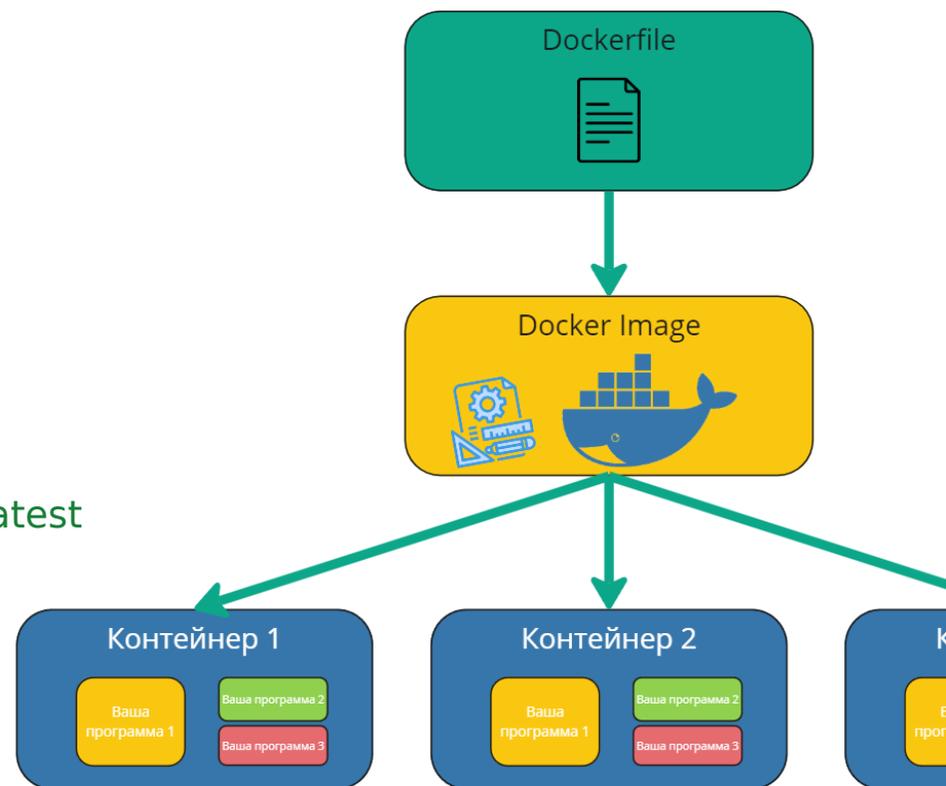
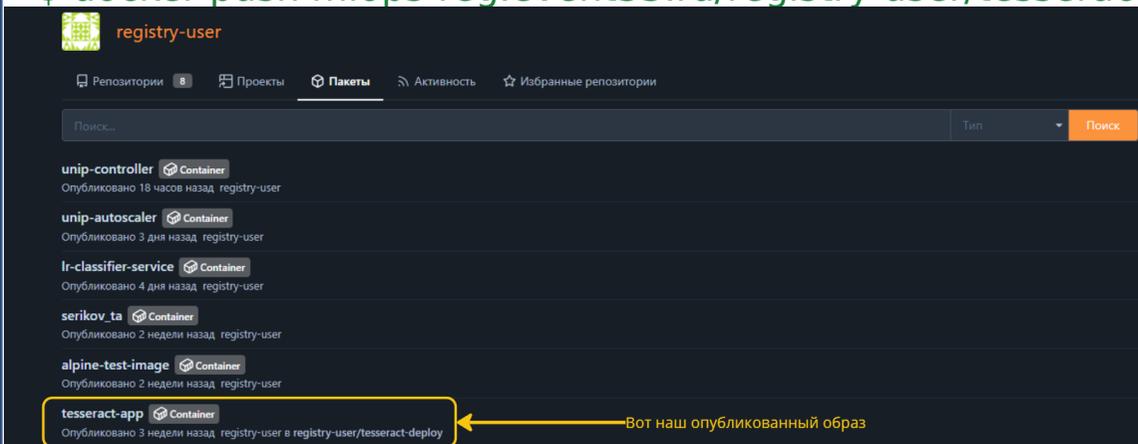
```
$ docker build -f service.dockerfile
```

Войдем в докер-репозиторий:

```
$ docker login mlops-reg.event33.ru
```

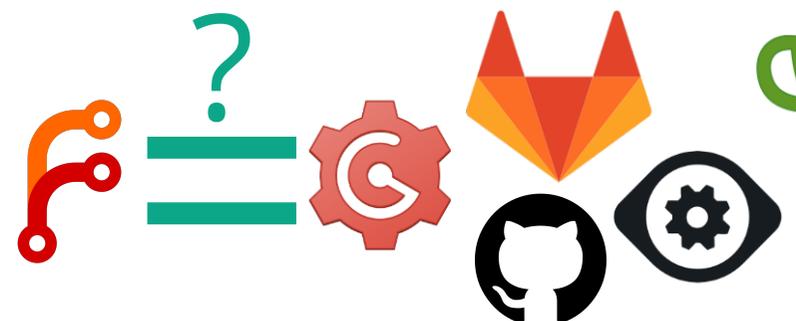
И запустим туда наш новый образ:

```
$ docker push mlops-reg.event33.ru/registry-user/tesseract-app:latest
```



4 Отправка репозитория на forgejo

- Forgejo — это бесплатная и открытая платформа для хостинга репозиторияев Git. Все знакомые нам git команды для работы с удаленными репозиториями будут работать с репозиториями на forgejo.
- Финальную версию проекта, подлежащего развертыванию нужно загрузить на ветку “deployment”.
- Сами файлы модели загружать на forgejo не нужно, их мы загрузим в S3 хранилище.



Структура git репозитория

