Определение категории товара для электронной доски объявлений

Люборт Константин Сергеевич
научный руководитель: Долбилов Владислав Георгиевич (Yandex)

НИУ ВШЭ СПб

8 июня 2021 г.

Предметная область

- Товары на досках объявлений организованы в иерархическую структуру
- Объявления размещаются в листьях дерева категорий
- Промежуточные узлы дерева используются для навигации при поиске товаров

Мотивация

- Количество узлов в дереве категорий исчисляется тысячами
- Пользователю сложно найти необходимую категорию в дереве
- Сложность выбора категории при размещении влияет не только на удовлетворенность продавца, но и на качество сервиса для покупателей

Существующие решения

- Использование заголовка объявления в качестве основного признака
- Группировка похожих категорий:
- eBay Research Labs[1]: Два уровня классификации. Выделение групп классов с помощью собственного алгоритма. KNN + SVM.
- Alibaba Group[2]: специальная функция потерь, учитвающая иерархию классов.

[1] Shen, Ruvini и Sarwar, “Large-scale item categorization for e-commerce”, 2012.

[2] Gao и др., “Deep Hierarchical Classification for Category Prediction in E-commerce System”, 2020.

Существующие решения

- Несколько решений используют дополнительные текстовые признаки
- Naver Shopping[3]: Текстовые признаки, специфичные для сервиса. Отдельные RNN для каждого признака.
- Rakuten Ichiba[4]: Заголовок и описание товара. Отдельные классификаторы для каждого признака.

[3] Ha, Pyo и Kim, “Large-scale item categorization in e-commerce using multiple recurrent neural networks”, 2016.

[4] Cevahir и Murakami, “Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant”, 2016.

Яндекс.Объявления

- Электронная доска объявлений
  - Несколько тысяч категорий-листьев
  - Частые изменения в дереве категорий
- Необходимо определять самые вероятные категории объявления, чтобы упростить процесс размещения для пользователей

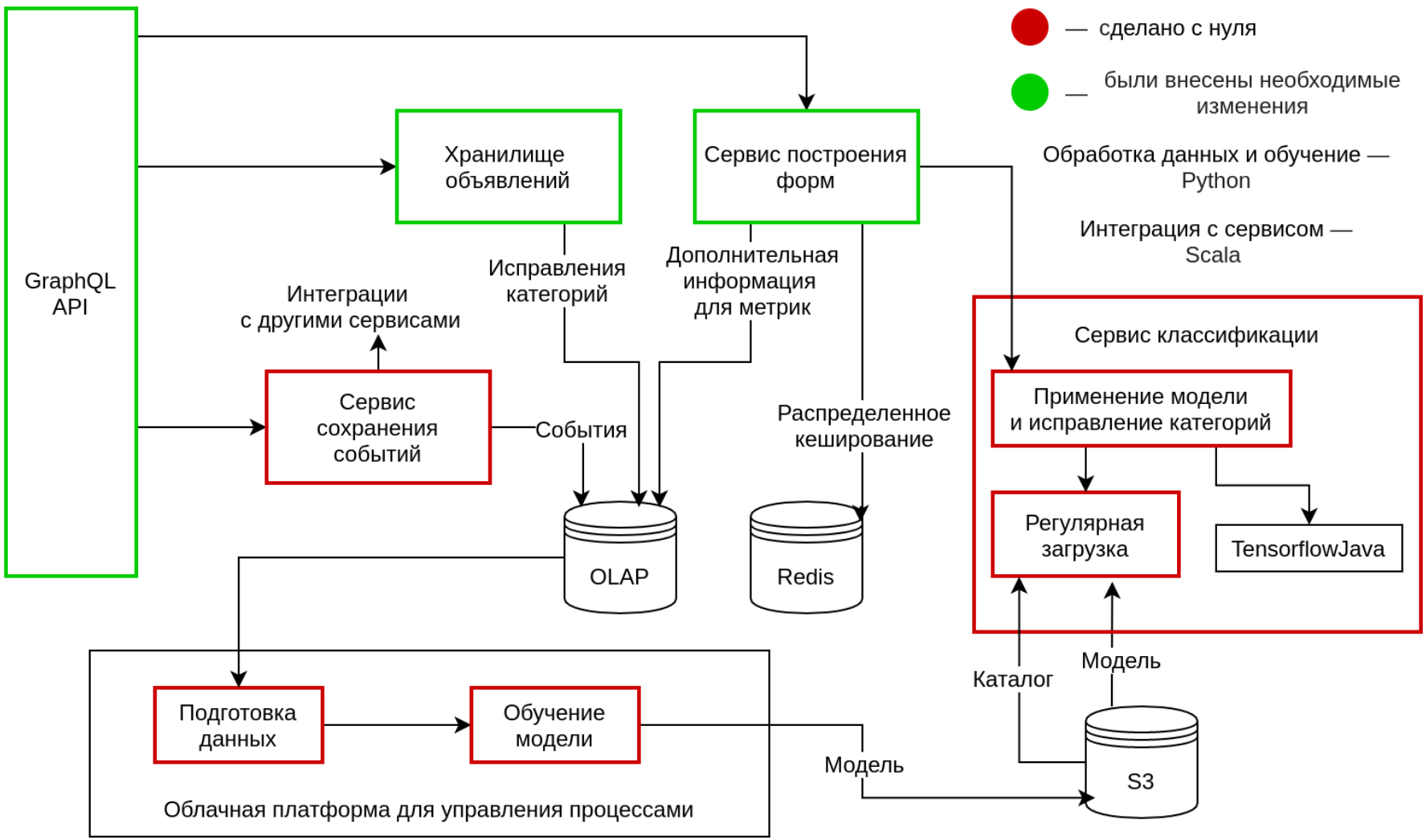
Цель и задачи

Цель: Разработать систему, помогающую пользователям выбрать категорию при размещении объявлений на сервисе Яндекс.Объявления

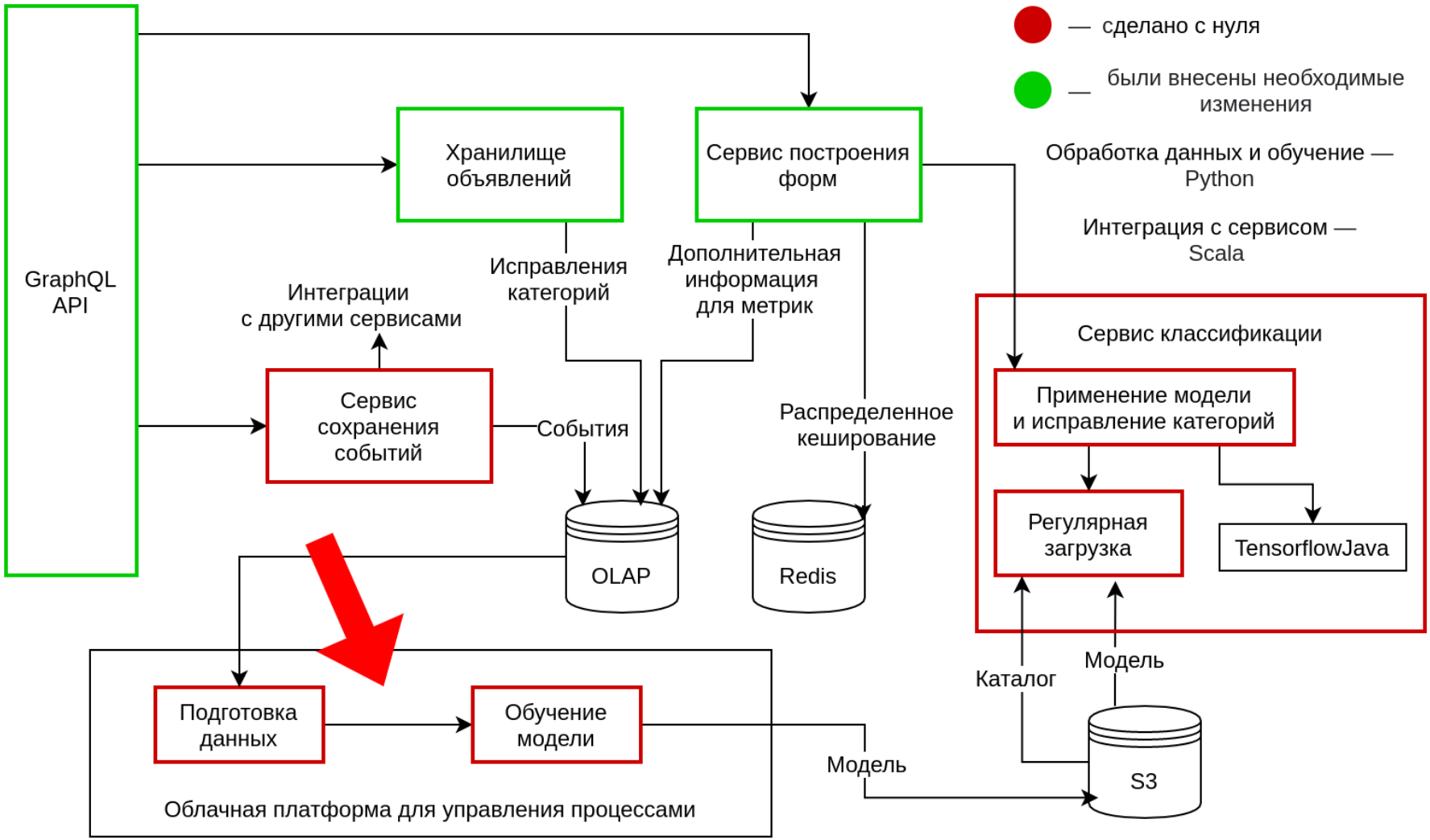
Задачи:

- Интегрировать предсказание категории в процесс добавления объявления
- Сравнить различные модели классификации, которые применяются для решения задачи
- Протестировать полученное решение на реальных пользователях

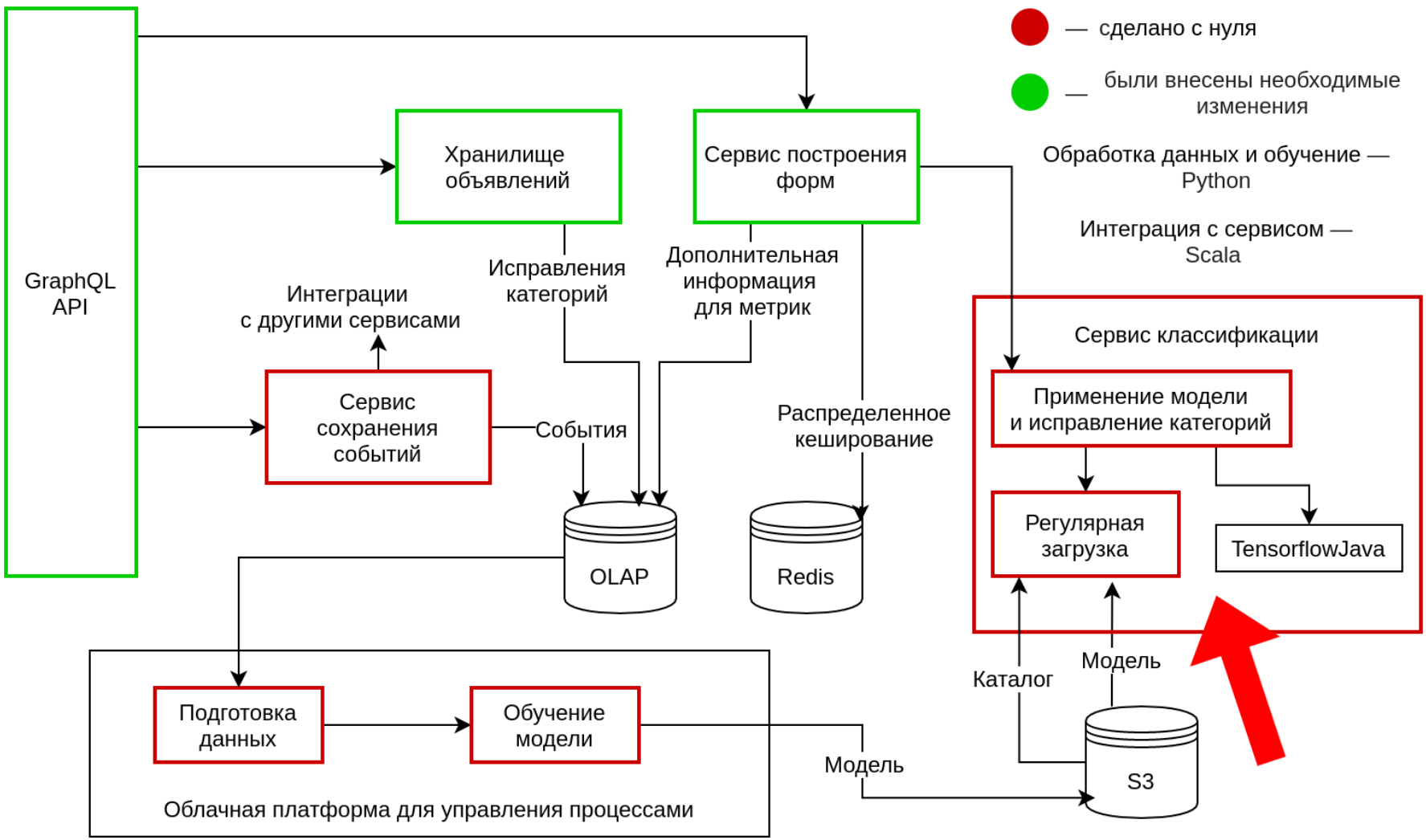
Архитектура



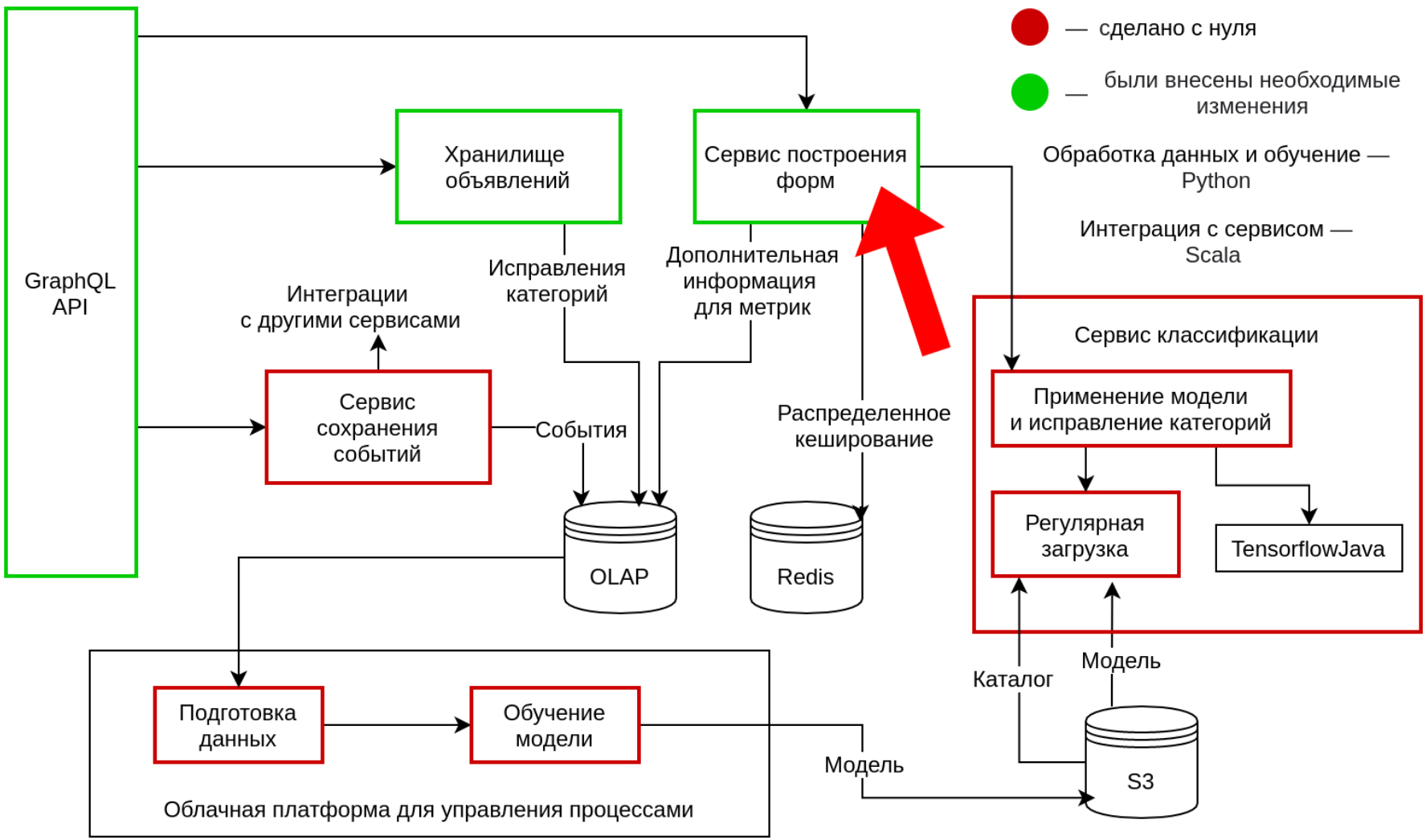
Архитектура



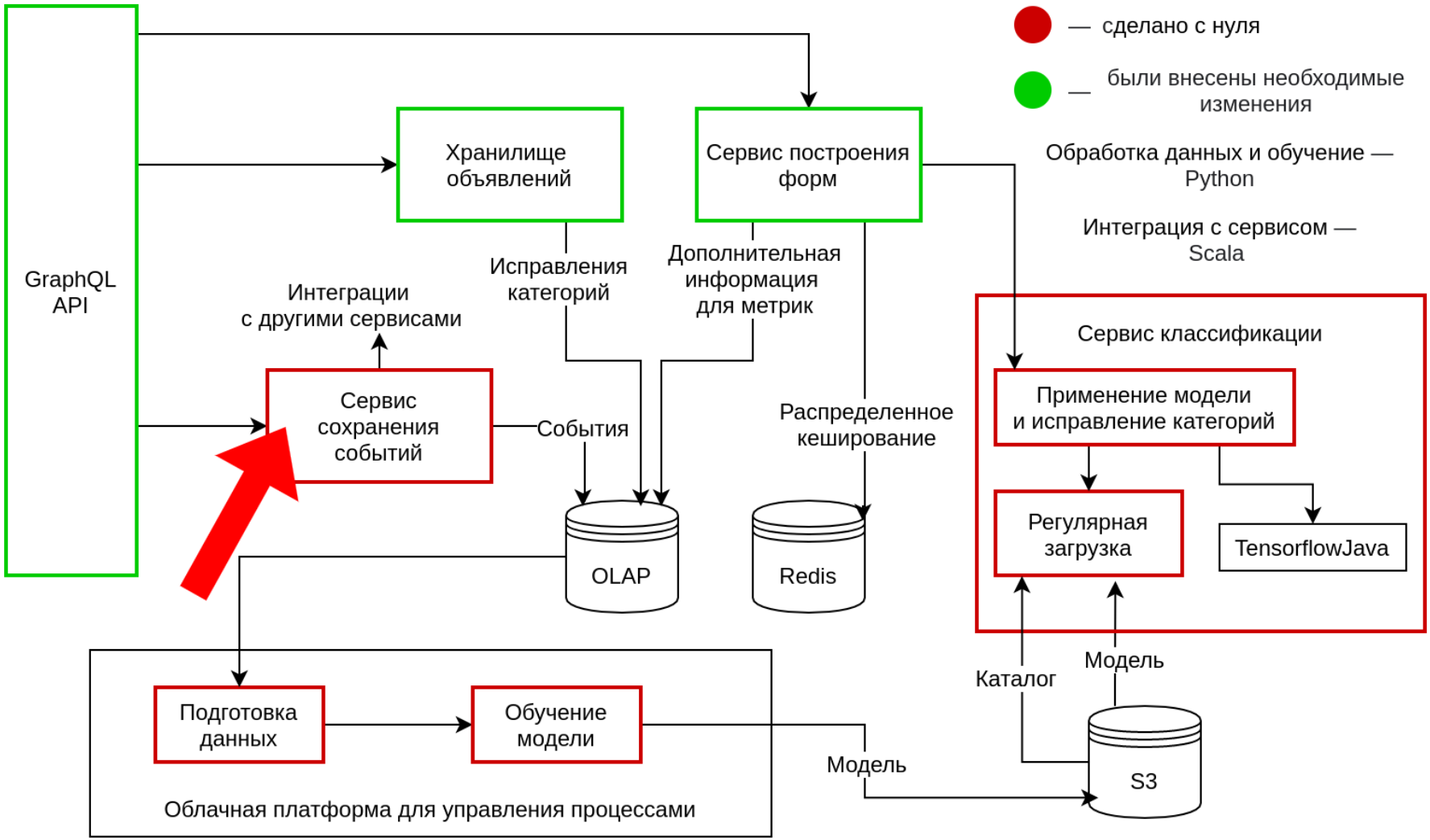
Архитектура



Архитектура



Архитектура



Датасет

- Обучающая (800k) и валидационная (200k) части:
  - Объявления, размещенные пользователями
  - Двойной вес для объявлений, в которых категория исправлена модераторами (+0.9% accuracy)
  - Примеры товаров из каталога, добавляемые сотрудниками сервиса (-0.15% accuracy; необходимо для новых категорий)
- Тестировочная часть (20k): разметка подмножества объявлений в Яндекс.Толоке. Разметка производилась сотрудниками сервиса.

Модель классификации

Модель	Accuracy	Top-3 Acc.	Top-5 Acc.
Multinomial NB	67.6	84.0	88.8
Pretrained DistilBERT	71.8	**89.3**	**93.6**
KNN-SVM	67.8	84.5	88.6
KNN-SVM + Кластеризация	69.1	85.2	89.5
DistilBERT + SVM + Кластеризация	71.5	87.7	92.0
DistilBERT + Alibaba (Кластеризация)	71.7	89.0	93.3
DistilBERT + Alibaba (Каталог)	**71.9**	88.7	93.0

- Имплементировано решение из статьи Ebay
- Получилось улучшить алгоритм из статьи заменой одного шага на кластеризацию
- С заменой KNN на нейронную сеть работает все равно хуже бейзлайна

Модель классификации

Модель	Accuracy	Top-3 Acc.	Top-5 Acc.
Multinomial NB	67.6	84.0	88.8
Pretrained DistilBERT	71.8	**89.3**	**93.6**
KNN-SVM	67.8	84.5	88.6
KNN-SVM + Кластеризация	69.1	85.2	89.5
DistilBERT + SVM + Кластеризация	71.5	87.7	92.0
DistilBERT + Alibaba (Кластеризация)	71.7	89.0	93.3
DistilBERT + Alibaba (Каталог)	**71.9**	88.7	93.0

- Имлементировано решение из статьи Alibaba Group.
- В качестве иерархии классов тестировалась группировка похожих классов из статьи Ebay и иерархия из каталога

Модель классификации

Модель	Accuracy	Top-3 Acc.	Top-5 Acc.
Multinomial NB	67.6	84.0	88.8
Pretrained DistilBERT	71.8	**89.3**	**93.6**
KNN-SVM	67.8	84.5	88.6
KNN-SVM + Кластеризация	69.1	85.2	89.5
DistilBERT + SVM + Кластеризация	71.5	87.7	92.0
DistilBERT + Alibaba (Кластеризация)	71.7	89.0	93.3
DistilBERT + Alibaba (Каталог)	**71.9**	88.7	93.0

- Получается улучшить метрики решения с помощью использования описания в качестве дополнительного признака

Модель	Accuracy	Top-3 Acc.	Top-5 Acc.
DistilBERT + Описание	73.1	89.3	93.4

- С описанием модель работает в среднем на 70% дольше (110ms вместо 65ms)

Результаты

- Процесс классификации пользовательских объявлений интегрирован в работу сервиса
- Имплементировано и протестировано 8 возможных решений с разными параметрами
- Выбранное решение показывает 71.8% accuracy при offline-тестировании
- Модель раз в 3 дня обучается на новых данных.
- Собираются пользовательские метрики текущего решения. В 70.9 % случаев пользователям подходит первая предложенная категория. Вручную приходится выбирать категорию для 4.7% объявлений.

Список литературы

- Cevahir, Ali и Koji Murakami. “Large-scale Multi-class and Hierarchical Product Categorization for an E-commerce Giant”. B: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, с. 525—535.
- Gao, Dehong и др. “Deep Hierarchical Classification for Category Prediction in E-commerce System”. B: *arXiv preprint arXiv:2005.06692* (2020).
- Ha, Jung-Woo, Hyuna Pyo и Jeonghee Kim. “Large-scale item categorization in e-commerce using multiple recurrent neural networks”. B: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, с. 107—115.
- Shen, Dan, Jean-David Ruvini и Badrul Sarwar. “Large-scale item categorization for e-commerce”. B: *Proceedings of the 21st ACM international conference*