Интеллектуальный анализ работы хранилища данных на основании обработки логов

ТРЕК ОТ РОСТЕЛЕКОМ.

ЭКСПЕРТЫ - АНДРЕЙ ТЕЛЮКОВ, НИКИТА БОГДАНОВ.

КОМАНДА BBYBY - УЛЛУБИЙ КАГЕРМАНОВ, НИЗАМОВА ЛИЛИЯ, ВЕРЕТНОВ ТИМОФЕЙ

SKOLKOVO HACK
2022

Идея проекта:

- НАЙТИ ВЫСОКОНАГРУЖЕННЫЕ “ГОРЯЧИЕ” ОБЪЕКТЫ ХРАНИЛИЩА ДАННЫХ И СПРОГНОЗИРОВАТЬ ПОТЕНЦИАЛЬНЫЙ РОСТ НАГРУЗКИ.

SKOLKOVO HACK
2022

Решение проблемы:

- Разработали алгоритм быстрой обработки данных.
- Выделили отдельные объекты, к которым чаще всего обращаются пользователи, как самых “тяжёлых” и самых “горячих”.
- Смоделировали прогнозирование нагрузки на систему в разные временные отрезки.

SKOLKOVO HACK
2022

CJM пользователя продукта

1. Клиенты:
   - пользователи, дата инженеры, администраторы баз данных
2. Для пользователей
   - информация о приблизительном времени на запрос
3. Для инженеров и администраторов
   - информация о "горячих" и "тяжелых" объектах и выделение под них дополнительных ресурсов,
   - информация о периоде времени нагрузки на систему либо в процессе выполнения запросов либо в предсказательных целях

SKOLKOVO HACK
2022

CJM пользователя продукта

Как использовать:

- телеграм бот или встроенный к код python скрипт
- ввод данных и получение ответа:
  1. топ(n) "горячих" таблиц"
  2. топ(n) "тяжелых таблиц"
  3. имя юзера(loguser) и запрос с получением среднего времени его исполнения.

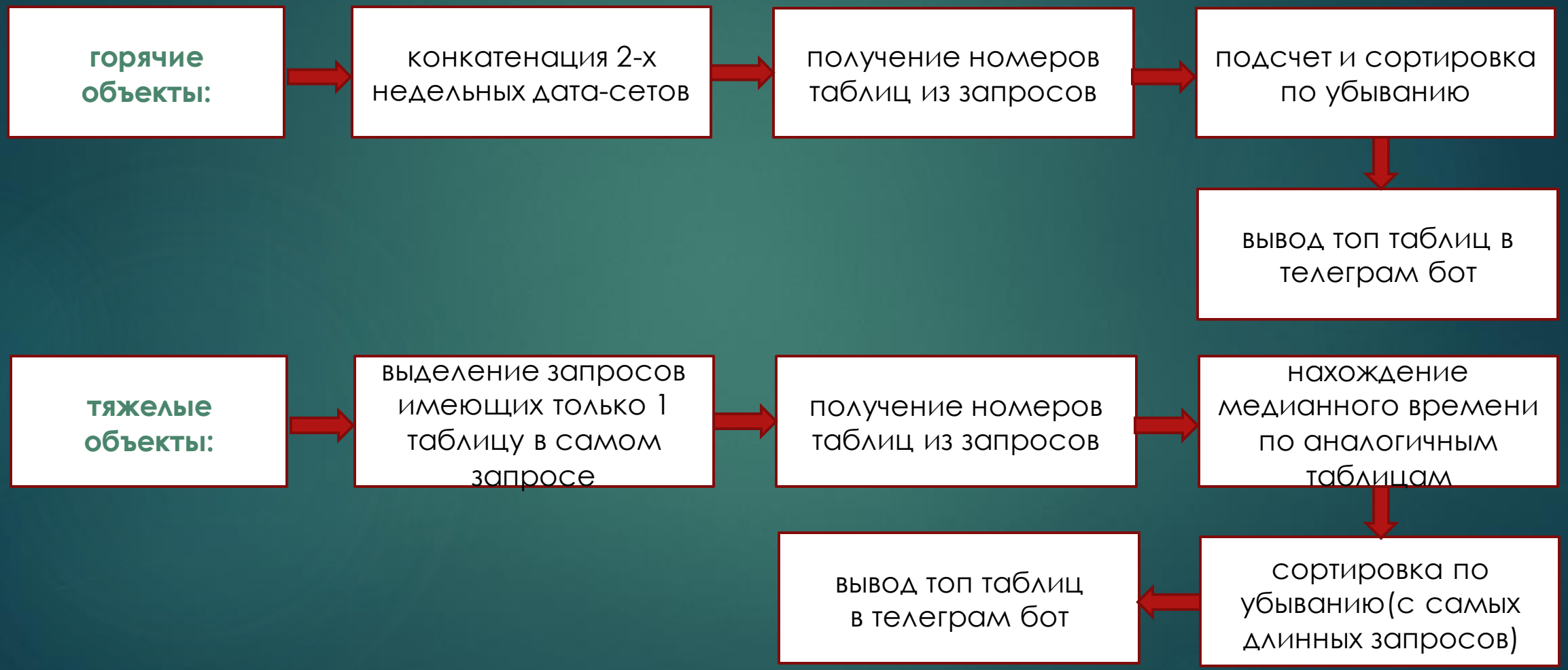
SKOLKOVO HACK
2022

Перечень используемых технологий

- Используемый стек: Python

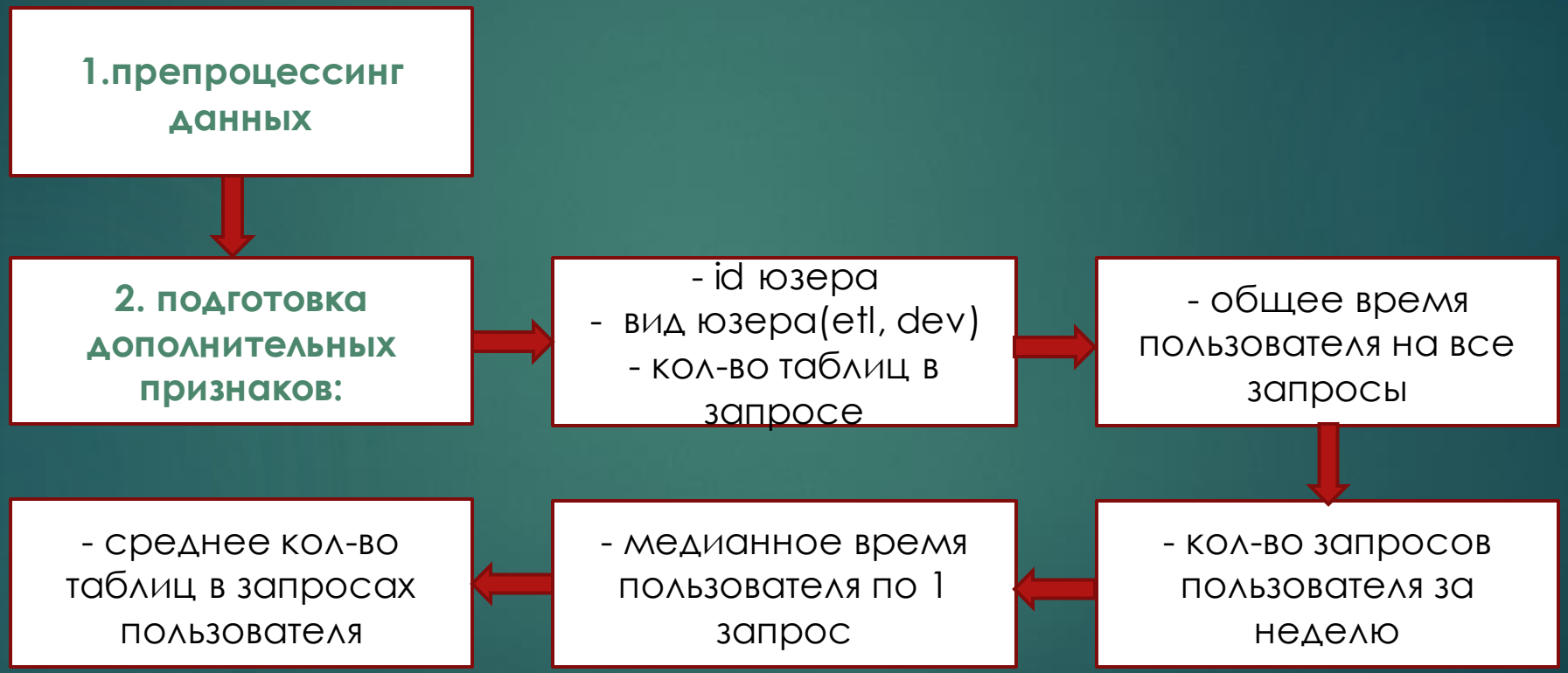
pandas, numpy, collections.Counter,
catboos (CatBoostRegressor, cv, Pool),
sklearn.model_selection (train_test_split), re,
sklearn.feature_extraction.text (CountVectorizer),
scipy.sparse, telebot, config.tokens

Технические характеристики решения



Технические характеристики решения

Прогнозирования времени исполнения запроса по предоставленным данным



Технические характеристики решения

Прогнозирования времени исполнения запроса по предоставленным данным

3. Удаление аномальных значений из данных (слишком большое время на запрос) → **4. Векторизация запросов**

5. Обучение данных (векторы + доп признаки) на модели (catboost). → Получение средней абсолютной ошибки на валидации в 1 сек.

6. Вывод среднего времени запроса по пользователю по запросу в телеграм боте

SKOLKOVO HACK
2022

Дополните-льная информация

Для отображения полученных нами данных мы используем телеграм бота

Функционал:

- оценка времени
- по шаблонным запросам
- по введёному запросу
- получения Топ Горячих и Тяжелый Таблиц
- расчет среднего времени выполнения запроса в зависимости от таблицы и пользователя

SKOLKOVO HACK
2022

Что не успели сделать

- оформить весь ресёрч из jupiter notebook в продакт версию с py файлами,
- связать телеграм бота с нашими алгоритмами и моделью

SKOLKOVO HACK
2022

Как можно было бы улучшить, имея дополнительные данные:

- улучшить точность прогнозирования времени запроса, имея полный запрос с агрегациями и сортировками.
- прогнозировать загруженность таблиц по дням, неделям и времени суток, имея timestamp
- прогнозировать загруженность нагрузки на систему(ЦМ, RAM, SDD), имя данные о нагрузках