sports.ru

BIG DATA В МЕДИА: ИНФРАСТРУКТУРА АНАЛИТИКИ НА OPEN SOURCE РЕШЕНИЯХ

Владислав Денисов

SPORTS.RU

Спортс – digital-издательство для неравнодушных к спорту

MAU

20 млн

уникальных пользователей

DAU

1,5 млн

уникальных пользователей

SOCIAL MEDIA

8 млн

уникальных пользователей

У нас огромное число UGC-контента, мы снимаем фильмы, создаем игры и много чего еще

здесь подробнее про нас

ДАННЫЕ В SPORTS.RU

Ключевые для бизнес-логики

Спортивная статистика
(~ 3k событий на матч)

Новости и материалы (~ 3k в месяц)

Активность пользователей:
комментарии (до 80k в сутки),
рейтингование (до 250k в сутки)
и подписки

Картинки
и другие медиа-файлы

ДАННЫЕ В SPORTS.RU

Аналитические

Кликстрим: ежедневно +50 ГБ сжатых данных о взаимодействии пользователей с сайтами, приложениями и другими сервисами

Экспорты из разных платформ (реклама, CRM, технический мониторинг и т.д.) для обогащения отчетов

КОМАНДА

- 4 «data fullstack» человека
- Нет разделения по ролям (веб / мобайл / продукт / маркетинг / бизнес / еще что-то)
- Есть условное разделение по зонам ответственности (описание, инжиниринг, nlp, исследования, спорт)
- Нанимали людей даже без знания SQL, главное – математика
- Аналитик – не калькулятор и не мастер по презентациям: все отчеты автоматизированы

АНАЛИТИКА

Базы данных

АНАЛИТИКА

ClickHouse

- Колоночная СУБД, изначально разработанная Яндексом (сейчас отдельная компания), хранилище для нашего кликстрима и других сервисов, где важны не единичные строки, а агрегаты.
- Используем с 2016-го года, тогда – удачная находка (скорость работы позволяла закрыть глаза на некоторые баги и непривычный синтаксис и бросать все силы на переезд с Redshift), сейчас – один из лидеров в области.
- Благодаря шардированию (параллельное вычисление и хранение данных) и репликации (дублирование данных между серверами для отказоустойчивости) можно не терять в скорости с увеличением объемов, данных, в отличие от классических версий MySQL / PostgreSQL.

АНАЛИТИКА

Инструменты

kubernetes

FastAPI

Apache Airflow

re dash

Streamlit

АНАЛИТИКА

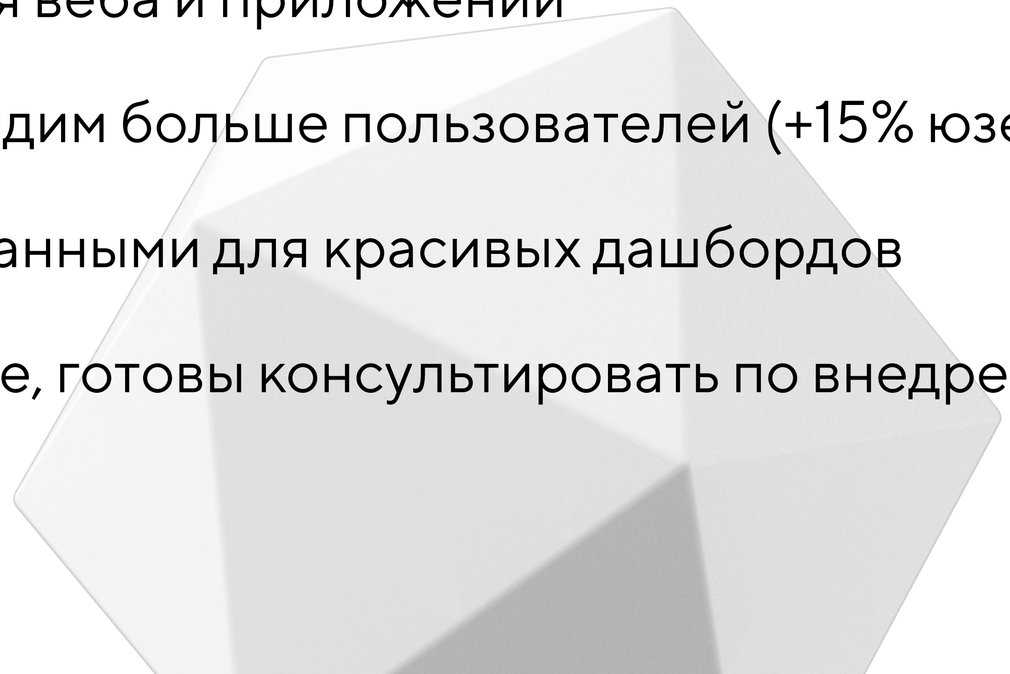
Инструменты

- Airflow – основной инструмент для создания пайплайнов по обработке данных. Благодаря этому сервису управлять всеми операциями стало проще – помогают встроенный алертинг и удобный интерфейс.
- FastAPI – основной инструмент для написания бекенд-сервисов в аналитике. Низкий порог входа позволяет создать эффективно работающую логику даже джунам с парой месяцев опыта.
- Kubernetes – система управления кластерами контейнеризированных приложений и сервисов. Де-факто стандарт в веб-разработке, который сильно упрощает развертывание и поддержку приложений.

АНАЛИТИКА

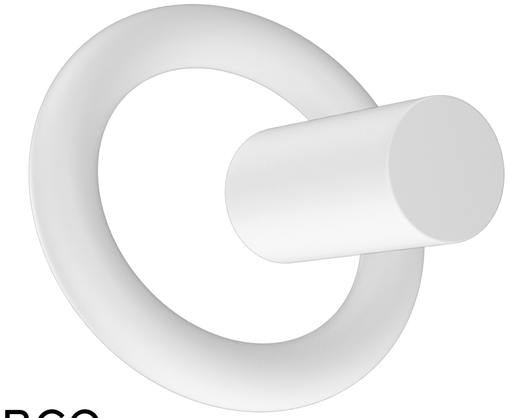
Свой счётчик. Зачем, когда есть GA/ЯМ?

- Возможность построения кастомных метрик и отчетов
- Мощная детализация (фильтры по конкретному пользователю или последовательности событий)
- Единая структура для веба и приложений
- Обходим адблок – видим больше пользователей (+15% юзеров)
- Соединение с метаданными для красивых дашбордов
- Открыт в open-source, готовы консультировать по внедрению / использованию



АНАЛИТИКА

Визуализация



- Минусы своего счетчика – на выходе просто таблица в БД, все представления данных нужно собирать с нуля
- BI – Redash (даже контрибьютим в развитие) – огромный набор визуализаций и возможность собирать дашборды на Python
- Если нужна интерактивность – пишем приложение на Streamlit (Python)

Fantasy 2.0

Refresh

Date

13/11/22 → 16/11/22

Tournament

world-cup

Source_type

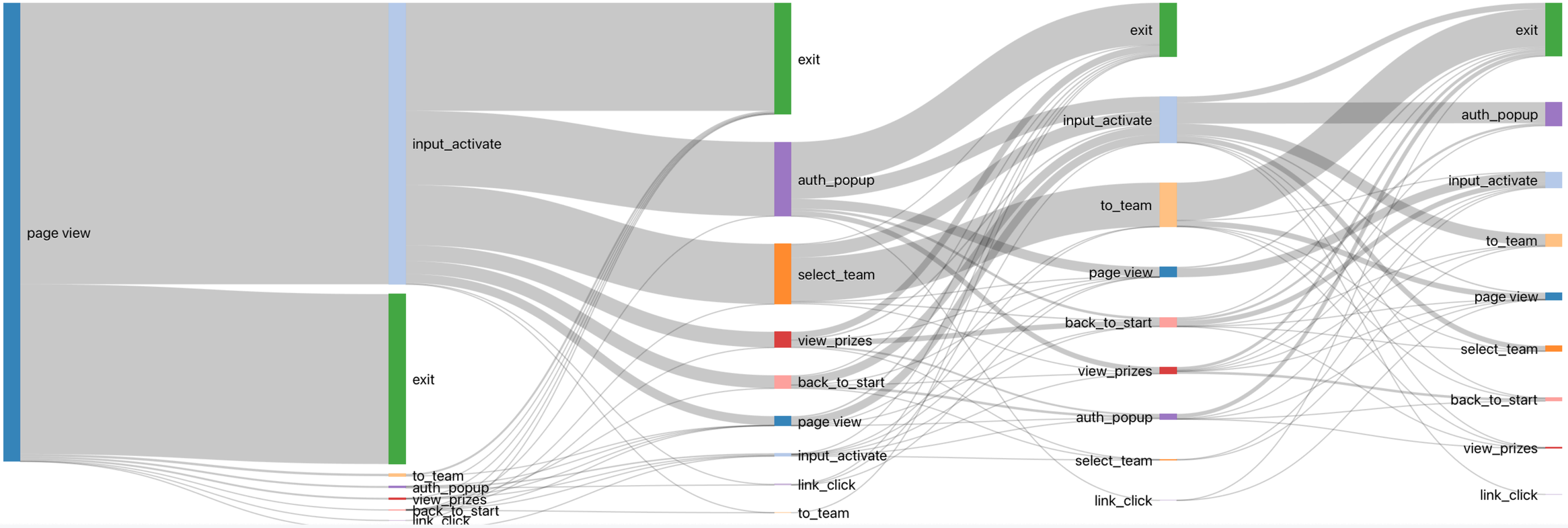
unknown

Fantasy | Onboarding | Sankey

Описание: https://docs.google.com/spreadsheets/d/1ZwiQReLj00TXefNgPt3YMVYXmMoHWGM3IZw1HUa-Tcc/edit#gid=2131333336&range=D141:D148

Tournament

world-cup



main

oracle

sync

☐ режим выгрузки

product | adblock | robot

sports | All | All

☐ исключить просмотры маркетинга

platform

mobile web ×

geo

Мск и МО ×

Прогноз

Чем больше период, тем точнее прогноз.

existing start | existing end

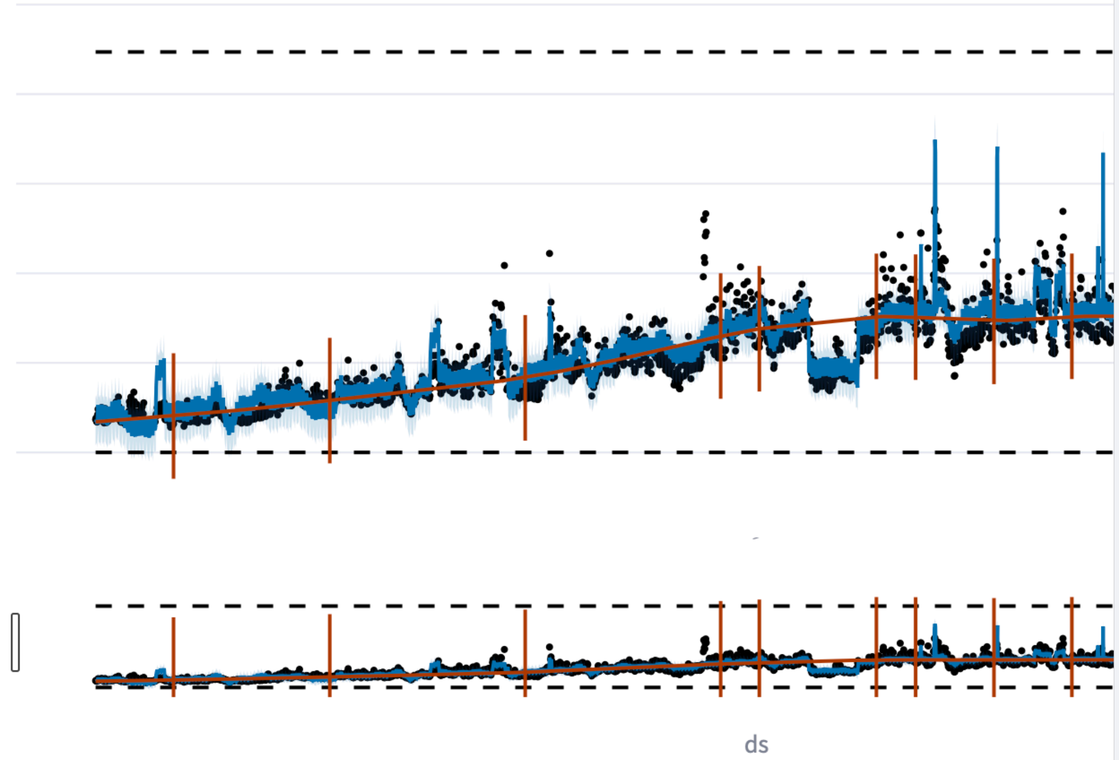
Advanced

Построить

Динамика

☐ Выгрузить в Google Sheets

1w 1m 6m 1y all



Аномалии

АНАЛИТИКА

Конкретнее про бигдату и детализацию

- Суммарно около 1000 строк в таблице с описанием событий
- Отслеживаем просмотр страницы и явные и неявные действия пользователей на них, каждое действие – 1 строка в базе, а 1 просмотр страницы генерирует около 100 событий
- Примеры вопросов, на которые умеем отвечать: какая дочитываемость у контента? а только у тех, кто заходит 5 дней подряд и чаще всего читает про Спартак? сколько денег нам приносит пользователь из ВК? какое пересечение у пользователей между вебом и приложениями? что мешает пользователям в процессе регистрации?

АНАЛИТИКА

Data Discovery

- Знания по структуре БД и событий не должны лежать в людях – выносим в общедоступные ресурсы и стараемся поддерживать актуальность
- Документация ивентов – в Google Sheets
- Документация БД – в DataHub

Google Sheets

DataHub

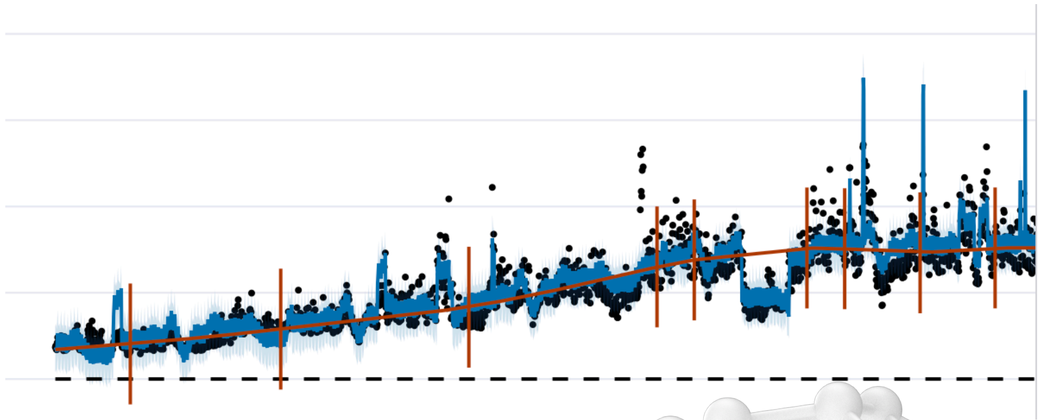
СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

- Аналитики в Sports.ru почти не занимаются ресерчем в ML: часто используем лучшие практики в уже готовых библиотеках
- Команда аналитики сама пишет backend для таких сервисов
- Нейросети – не панацея, если классические алгоритмы справляются не хуже, чаще «классика» работает значительно быстрее, что важно для продакшна
- Есть задачи, которые приходят от бизнеса, но значимая доля сервисов рождается из желания попробовать новый инструмент

СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

Прогноз трафика

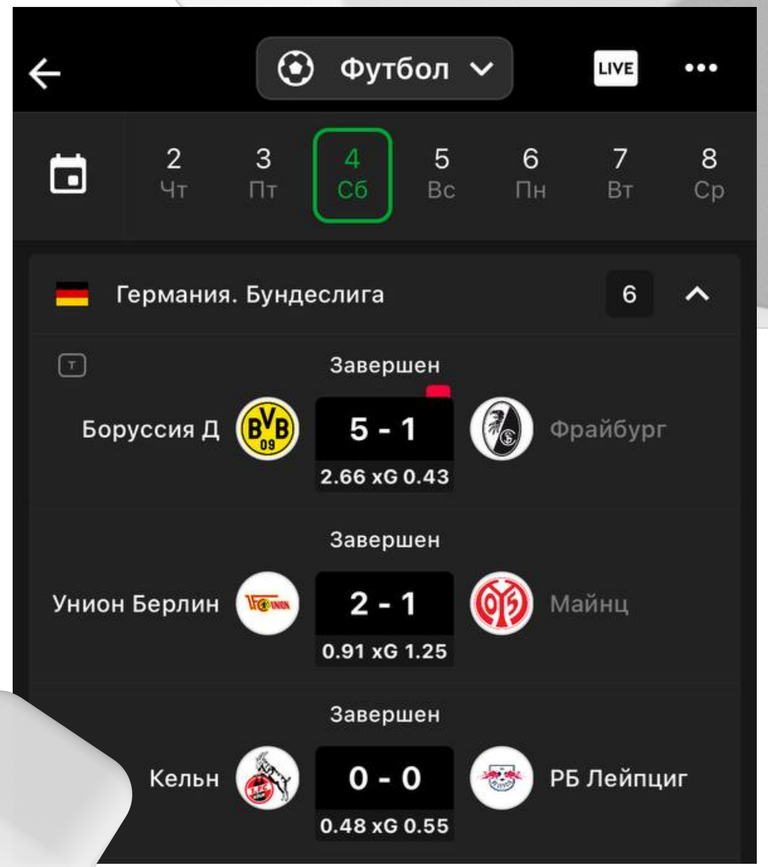
- У нас рекламная модель бизнеса – выручка коррелирует с просмотрами страниц
- Делаем краткосрочный (месяц) и долгосрочный (год) прогнозы
- Сервис прогноза – инструмент для отдела продаж – делаем так, чтобы рекламы было ровно столько, сколько инвентаря доступно



СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

Expected goals в футболе

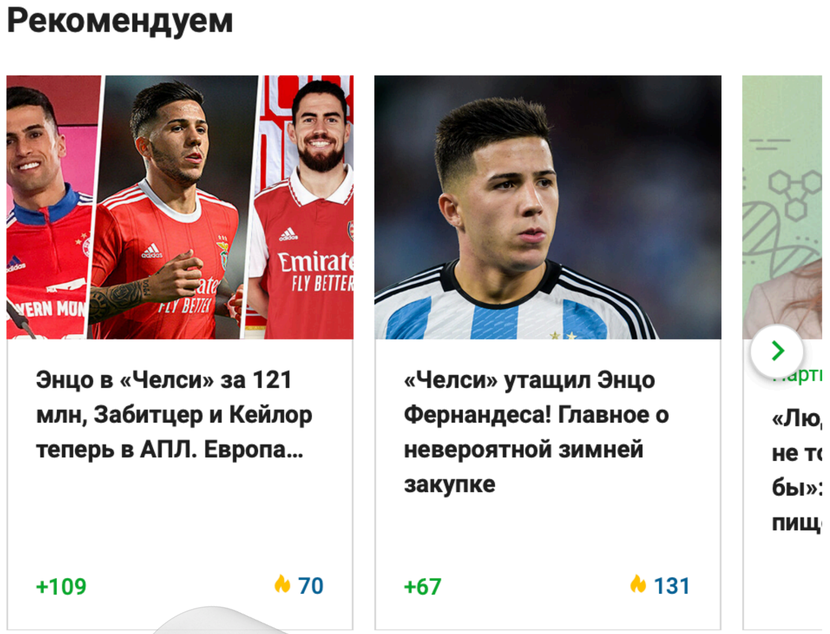
- xG – показатель расширенной футбольной статистики, который показывает остроту моментов
- Обучили модель классификации на основе бустинга, которая предсказывает вероятность гола по 100+ параметрам
- Топ-5 лиги + РПЛ имеют данный показатель, который обновляется прямо во время матча



СЕРВИСЫ НА ОСНОВЕ ДАННЫХ

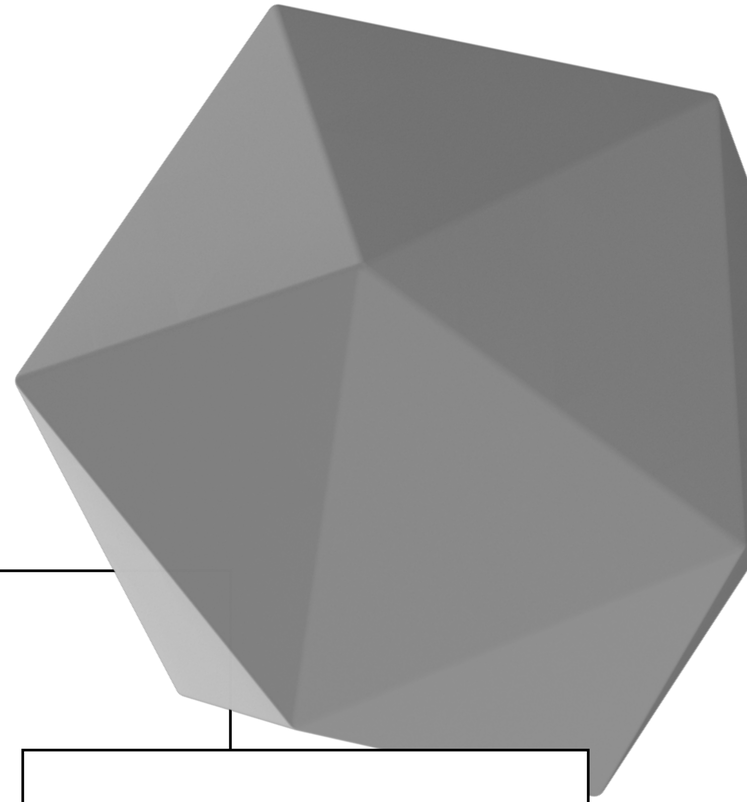
Кросспроектные рекомендации

- Делаем свою систему рекомендаций с 2017-го года, сейчас подключаем сторонние ресурсы (пока бесплатный SaaS)
- Работаем качественнее «обменок» (CTR + глубина), но без кликбейта
- Множество алгоритмов: многорукие бандиты, «классика», item2item, next-item – подстраиваемся под конкретную задачу



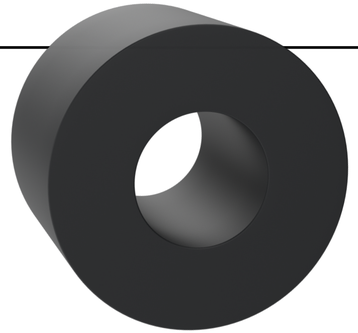
ВСЕ!

Остались вопросы? Хотите потестировать рекомендалку или счетчик аналитики? Пишите!



TELEGRAM: @VLADENISOV

MAIL: DENISOV@SPORTS.RU



sports.ru