Кликстрим в Авито

Собираем. Обогащаем. Анализируем.

Дмитрий Хасанов

Обо мне

- фриланс
- классифайды
  n1.ru, auto.ngs.ru, avito.ru
- хакатоны
  Hack. Sleep. Repeat: goo.gl/bzMjjC

Авито

- топ 5 сайтов России по посещаемости
- аудитория 35 млн в месяц
- третья по стоимости компания Рунета
- мноооожество объявлений
- не только Авито

План

- Что может аналитика
- Собираем
- Обогащаем
- Анализируем
- Предоставляем интерфейсы
- Обучаем пользователей
- Преодолеваем трудности
- Будущее системы

Что может аналитика

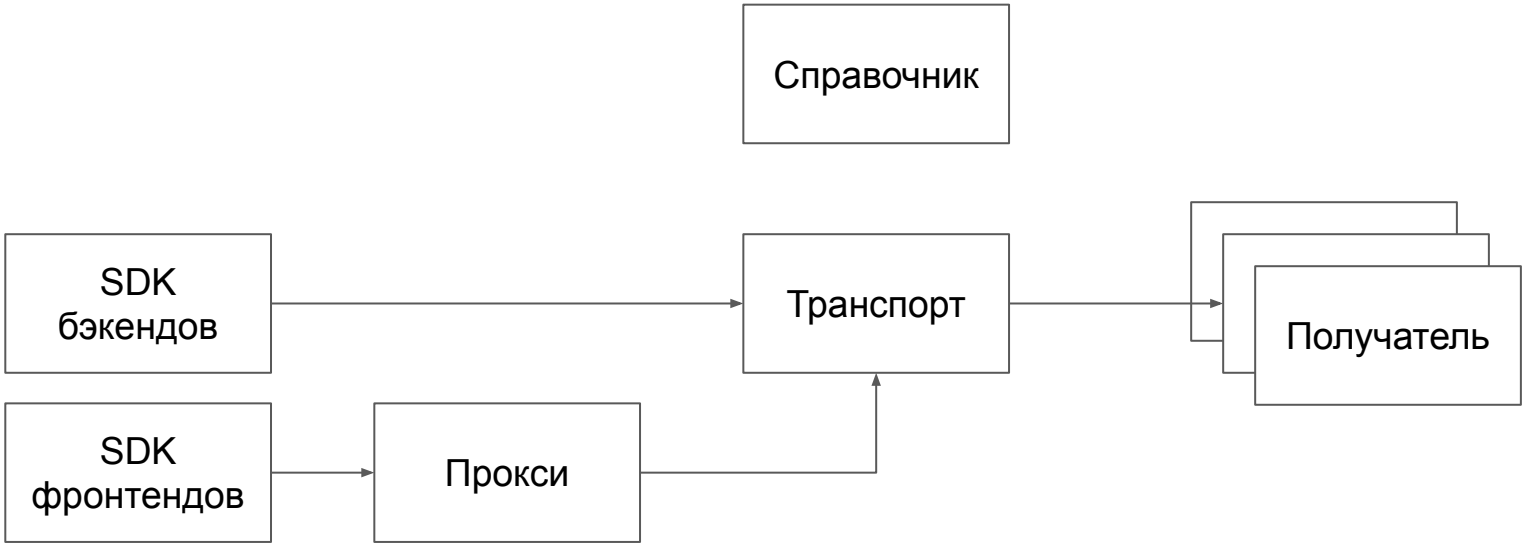
Первичные данные

- Факты
- Агрегаты
- Последовательность действий

Использование данных

- Данные для пользователей
- Отчёты, бизнес-модель
- Ad-hoc аналитика
- A/Б-тесты
  - Интерфейсные изменения
  - Тюнинг бизнес-модели
- Обучение ML-моделей
  - Качество поиска
  - Рекомендации
  - Антифрод

Кликстрим Авито



Собираем

Справочник

- Поля
- События
- Окружения
- Правила доставки получателям
- Владельцы событий
- Владельцы получателей
- Правила раскладки в хранилища

SDK

- Протокол
- Формирование и отправка событий
- Кодогенерация: go, php, python, js, android (kotlin), ios (swift)
- Валидация
- Буферизация, повторная отправка
- Дискавери транспорта

Транспорт

- NSQ
- Очень разные получатели
  - NSQ
  - Монго
  - Rabbit
  - Clickhouse
  - Statsd
- Динамические правила доставки

Прокси для фронтендов

- Прозрачный
- Защищённый
- Единственный

Хранилища

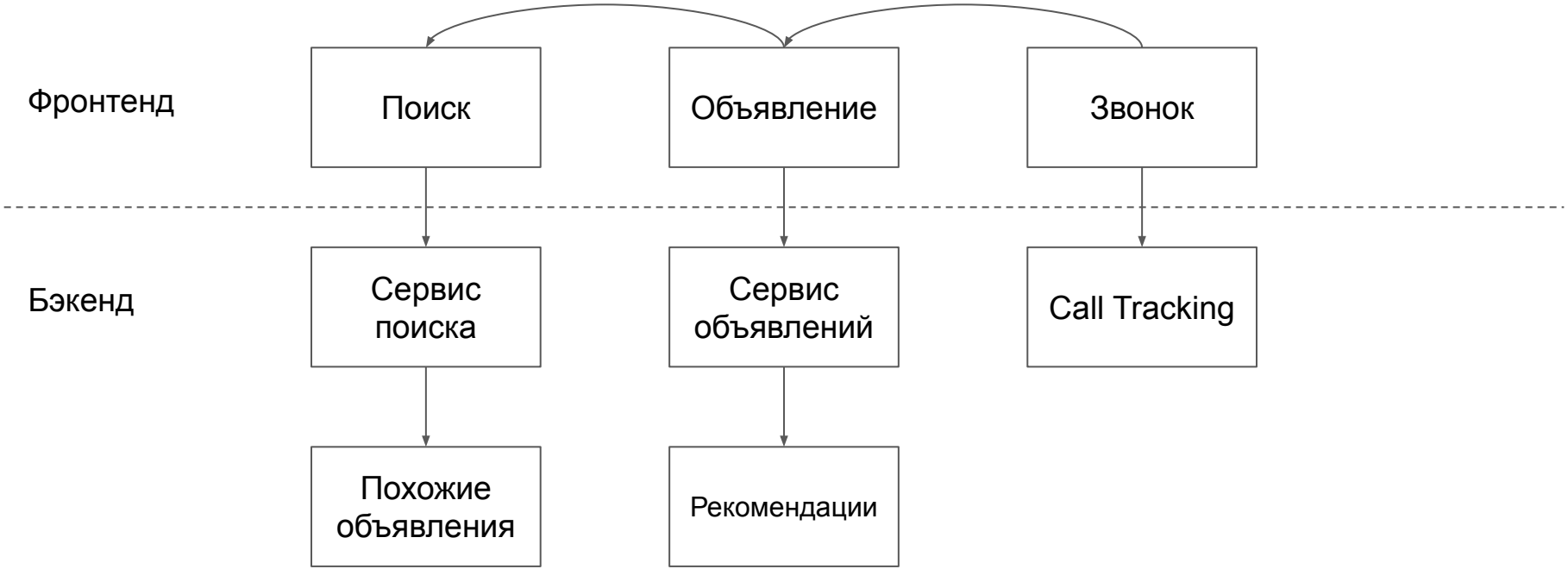
- DWH
  - Историчность
  - Нормализация
  - Общее назначение
  - Далеко от реалтайма
- Clickhouse
  - Ближе к реалтайму
  - Денормализация
  - Заранее определённый круг задач

Обогащаем

Необходимость обогащения

- Неполнота данных от фронтендов
- Сервисная архитектура
- Данные из внешних источников

Связь фактов



Точки обогащения

- Перед DWH
- Перед Кликхаусом
- В транспорте
  - На потоке
  - В буферном хранилище

Способы обогащения

- Поход в сервис
  - Синхронный
  - Асинхронный
- По словарю
- Объединение фактов

Анализируем

Отчёты

- Бизнес-модель
- Vertica + Tableau
- Историчность
- Воронки

Ad-hoc

- Vertica + Tableau
  - Полные данные
  - Историчность
  - Далеко от реалтайма
- Кликхаус
  - Неполные данные
  - Малый период хранения
  - Близко к реалтайму

А/Б-тесты

- Общий центр
- Маркеры тестов в событиях
- Нужна оперативность

Предоставляем интерфейсы

Документация

- Подключение кликстрима
- Проверка отправляемых событий
- Управление получателем
- Использование мониторинга

SDK

- go, python, php, js, android (kotlin), ios (swift)
- README.md
- Отдельный отправщик
- Репозитории пакетов: satis, pypi, gl

CMS справочника

- Поля
- События
- Окружения
- Консьюмеры

Мониторинг

- Событие
- Окружение
- Юнит
- Интеграция в CMS
- Канареечные релизы

Проверка отправки

- Сервис Eye of Providence
- Фильтры для отлова
- Интерфейс для просмотра
- Источники:
  - дев-среды
  - тестовые сборки
  - прод

Обучаем пользователей

Обучаем пользователей

- Документация
- Внутренние митапы
- Демо
- Синхронизационные встречи
- Совместный запуск продуктов

Преодолеваем трудности

Наследие

- монолит
- отправка событий без SDK
- дополнительные прокси

- информирование
- “продажа” новых подходов
- административный ресурс
- профилактика новых бед
  - мониторинг
  - отметки deprecated
  - документация
  - информирование
- grep
- много ручного труда

Хайлоад

- событий много
- узлы отказывают

- выбрать нужную надёжность
- убрать единые точки отказа
- масштабировать хранилища
- масштабировать буферы
- выбрать быстрый транспорт
- оставить возможности восстановления
- мониторинг, алертинг

Обратное давление (backpressure)

- один из получателей не справляется

- установить политику работы с получателями
- мониторинг и алертинг — владельцам получателя
- транспорт должен жить
- буферы для возможности восстановления

Разные языки

- go
- php
- python
- js
- kotlin
- swift
- lua?

- общий протокол
- помощь продуктовых команд
- документация на создание SDK
- тестовый стенд

Разные доменные области

- разные проекты
- похожие сущности
- поля могут быть одинаковыми

- универсальный протокол
- транспорт не содержит доменных знаний
- ETL не смешивает события разных доменных областей

Боты

- много событий
- смазывают картину

- разметка на потоке
  - статические признаки
  - эвристики
- очистка в DWH
  - статические признаки
  - эвристики
  - поведенческий анализ

Аномалии

- нетипичное изменение потока событий
  - всего потока
  - отдельных проектов
  - отдельных событий
- достаточная детализация метрик
- алерты на весь поток событий — команде кликстрима
- алерты на события и проекты — командам-владельцам

За кадром

За кадром

- Гарантии доставки
  - At most once
  - At least once
  - Exact once
- Проверки надёжности
  - End-2-end тесты
  - Трейсинг между узлами
- Время возникновения и время обработки события
- Реалтайм-аналитика

Будущее системы

Будущее

- Применимость реалтайм-аналитики
- Концепция фактов
- Проактивный сбор фактов

Материалы

Кликстрим в Авито: https://habr.com/en/company/avito/blog/419651/
Устройство DWH: https://habr.com/en/company/avito/blog/322510/

Спасибо

pik4ez@gmail.com
github.com/pik4ez