

МАТЕ

9-10 НОЯБРЯ / 2023

МАРКЕТИНГ

Г

2023

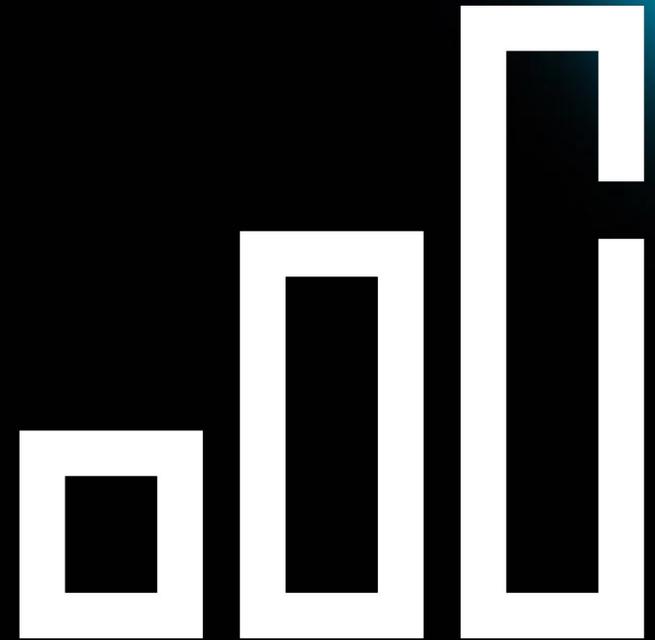
📍 GRAND BALLROOM / МОСКВА

БОЛЬШАЯ КОНФЕРЕНЦИЯ
ПО МАРКЕТИНГОВОЙ
И ПРОДУКТОВОЙ АНАЛИТИКЕ



БИЗНЕС-АНАЛИТИКА. КАК АНАЛИЗ ETL- ПРОЦЕССОВ ПОМОГАЕТ ОПТИМИЗИРОВАТЬ ХРАНИЛИЩЕ

АЛЕКСАНДР КИСЕЛЁВ,
руководитель бизнес-аналитики
медиахолдинга RAMBLER&Co



RAMBLER&Co

крупнейший медиахолдинг России

ЛЕНТА.RU

газета.ru

Рамблер/

 ЧЕМПИОНАТ

СЕКРЕТ
ФИРМЫ

МОСКВА
ЛЕНТА

 ferra.ru

Quto.ru

motor

 LIVEJOURNAL

No

1
медиахолдинг
по объему ежемесячной
аудитории цифровых
ресурсов (Mediascope, 2023)

>40%

пользователей Рунета
ежемесячно читают
СМИ Rambler&Co
(Mediascope, 2023)

>1

единиц контента в год —
тексты и видео
млрд

Задача бизнеса



Цель

Быть самыми актуальными в медиапространстве

Задача

Оперативное отслеживание конкурентов – практически в live-режиме.

Создание мониторинга событий.

Вопросы

Что у них сейчас собирает просмотры, что нам срочно нужно написать?

Сколько новость собрала просмотров после публикации?



Задача бизнеса



Идеальный сценарий – ежеминутное обновление информации по всем новостям, опубликованным за последние 7 суток:

~450 новостей в сутки (в среднем по одному ресурсу) *
7 суток = 3150 страниц

3150 страниц * 60 минут * 24 часов
= 4.5 млн запросов в день, или 58 запросов в секунду

5 сайтов * 4.5 млн = ~22 млн записей в день

Проблемы:

- недостаточная скорость получения данных;
- нагрузка на сайты;
- нагрузка на ресурсы (Airflow, сервер, БД);
- возникновение сложностей с оптимизацией работы скриптов / real-time аналитикой;
- просадки производительности.



Ограничения на количество запросов



Отсутствие динамически изменяемых прокси

- ~30 запросов в минуту на каждый сайт (консервативный темп);
- не требует отдельной реализации;
- можно столкнуться с блокировкой по IP.

Прокси

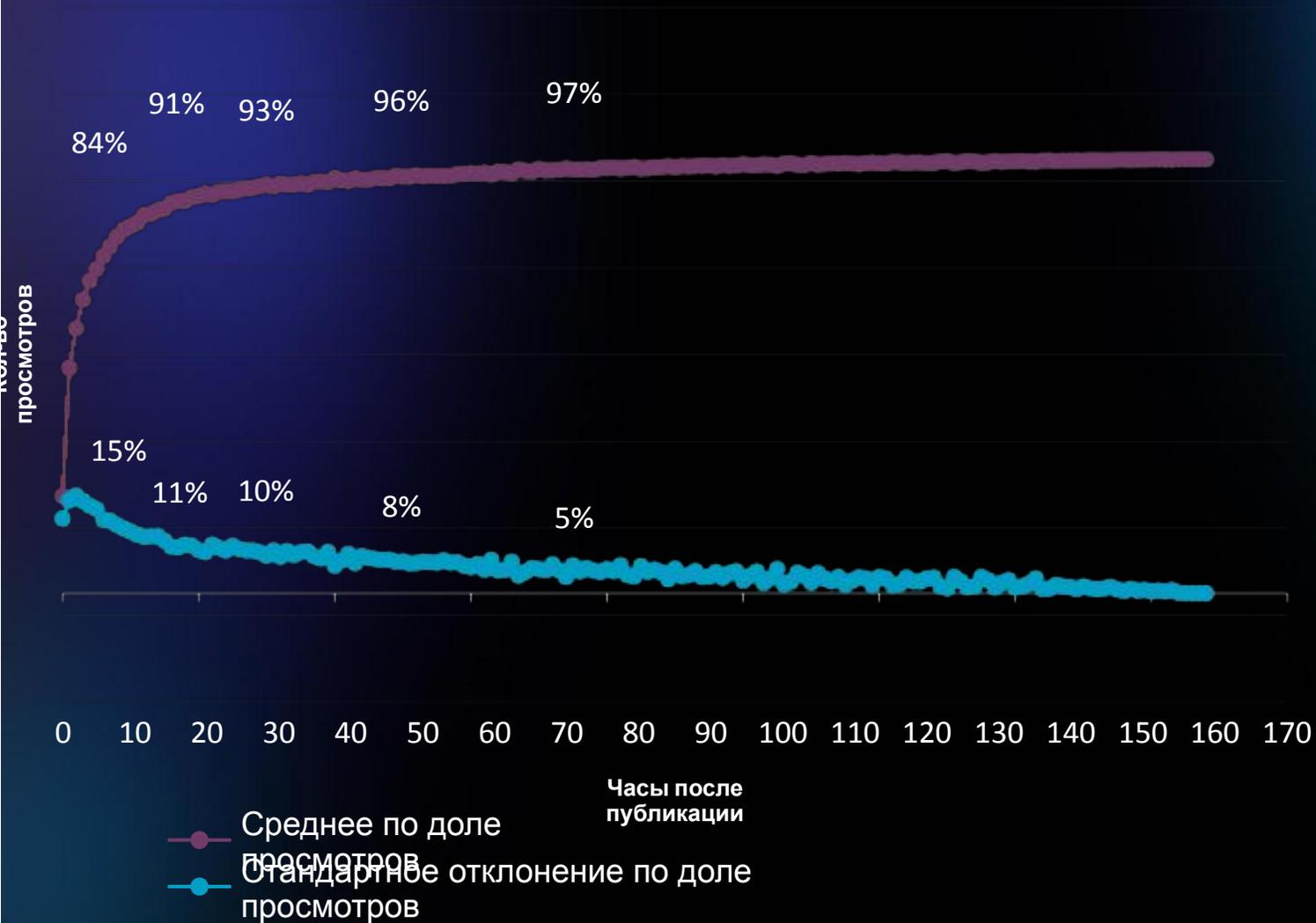
Бесплатные (при использовании напрямую):

- в ~98% случаев недоступные / медленные;
- сложный процесс отправки запросов;
- отсутствие значительного ускорения;
- снижение рисков блокировки.

Платные:

- на рынке представлено достаточное количество сервисов;
- в теории ускорение пропорционально количеству;
- небольшое усложнение процесса отправки запросов.

Распределение новостных событий



В первые 24 часа с момента публикации новость достигает пика набора просмотров (более 90% от всех просмотров)

Отношение объема показов к первым суткам

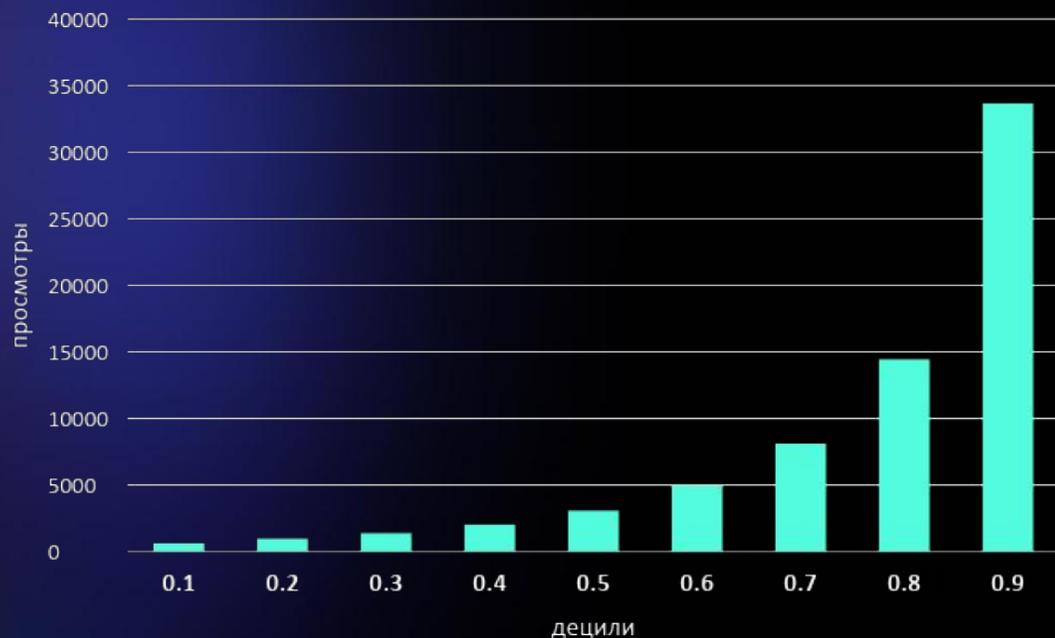
| | Среднее отношение | Стандартное отклонение | Количество наблюдений | Квантиль 9 |
|-----------------|-------------------|------------------------|-----------------------|------------|
| Вторые сутки | 4% | 10% | 1680 | 7% |
| Третьи сутки | 6% | 11% | 439 | 12% |
| Четвертые сутки | 4% | 5% | 151 | 9% |
| Пятые сутки | 6% | 7% | 102 | 15% |



Постановка ограничений



Количество просмотров



| ТОП рангов | Среднее | Медиана | Минимум |
|------------|---------|---------|---------|
| 1 | 181 687 | 167 091 | 25 210 |
| 2 | 124 404 | 120 052 | 8 583 |
| 3 | 100 409 | 100 745 | 5 945 |
| 4 | 85 806 | 87 816 | 1 604 |
| 5 | 75 124 | 77 570 | 1 272 |
| 10 | 53 820 | 53 166 | 5 026 |
| 15 | 40 090 | 42 599 | 3 178 |
| 20 | 31 097 | 32 925 | 1 858 |
| 30 | 20 736 | 22 750 | 728 |
| 40 | 15 537 | 16 742 | 379 |
| 50 | 12 334 | 13 541 | 177 |

- Статья на вторые сутки после публикации не может набрать просмотров более, чем за первые сутки
- 80% процентов статей набирают < 15 000 показов в течение первых суток после публикации
- Статьи не попадут в топ-40 статей во все последующие сутки
- Требуется обновлять просмотры только для 20% новостей от общего объема исторических данных

Решение: Организация процесса



Алгоритм парсинга

1. Тексты по новым статьям – высокая частота (новости опубликованные «только что»)
2. Тексты и просмотры по недавно опубликованным статьям – средняя частота (статьи в первые 24 ч.)
3. Тексты и просмотры по популярным статьям прошлых дней – низкая частота (статьи >1 дн. и <7 дн.)
4. Просмотры по непопулярным статьям прошлых дней – минимальная частота (непопулярные статьи низкая частота (статьи >1 дн. и <7 дн.)

Вариант расчета

Текущая скорость: 1 запрос ~ 2.5 секунды

1. max в минуту = 4 новости (20 за 5 минут)
 2. max в сутки = ~600 новостей – $600 * 5 / 60 = 50$ минут → 50 минут раз в час * 24 / 60 = 20 часов
 3. $450 * 6 * 0.2 = 540$ новостей – $540 * 5 / 60 = 45$ минут → 45 минут раз в 8 часов / 8 * 24 / 60 = 2.25 часов
 4. $450 * 6 * 0.8 = 2160$ новостей – $2160 * 2.5 / 60 = 90$ минут → 90 минут раз в 24 часа / 24 * 24 / 60 = 1.5 часа
- Итого = 23.74 часа (укладывается в сутки)



Airflow

- оркестрация производителей (producers)
- установка расписания и приоритетов

RabbitMQ

- менеджмент приоритетов
- мониторинг активности / ошибок
- сохранение данных в случае ошибок



Обработчик запросов:

- лимитирование нагрузки
- установка задержки

↕

Конструктор запросов:

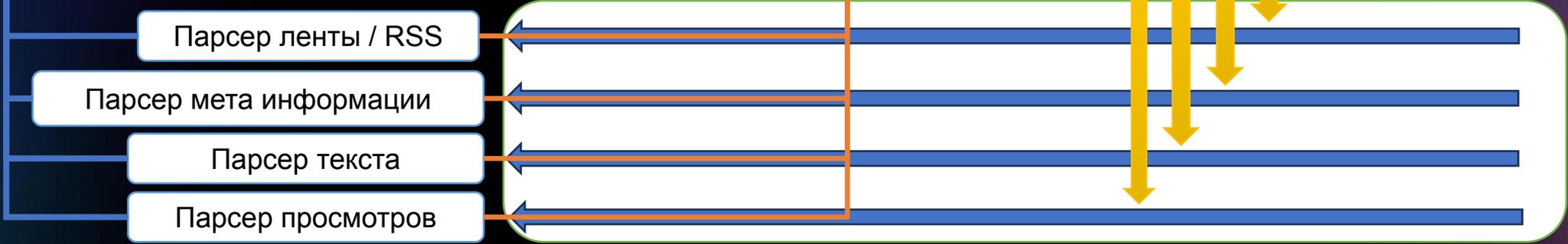
- установка headers
- пагинация

Система логирования

- Лента новостей, HTML
- RSS-feed, XML
- Новость, HTML
- Данные API, JSON

PostgreSQL

RabbitMQ





При поступлении задачи со стороны бизнеса необходимо:

- **Внедрение всестороннего анализа данных**, который поможет оптимизировать потребности без потери эффективности
- **Правильно сформированное ТЗ на реализацию, исходя из результатов анализа**, которое позволит организовать техническую схему реализации с оптимальной нагрузкой на вычислительные мощности

Предложенная схема поможет реализовать продукт, решая задачу бизнеса