

Практика перехода на ClickHouse и новые задачи для системы в развитии DMP

Развитие системы анализа медиапотребления

Денис Белослюдов

Директор Департамента
специальных проектов
«Агентство 2»



«Агентство 2»

Агентство 2 — рекламное агентство полного цикла

**Проведение
рекламных компаний**

Digital реклама

Рекламная сеть «А – Traffic»

**Продажа рекламы
в сети Триколор**

Медиаизмерения

Big Data «Триколор»



Проект «Медиаизмерения»

Исследования аудитории оператора цифровой среды «Триколор»

До 2022 года

Иностранное специализированное ПО для телеизмерений

После 2022 года

ПО «Аметрикс», на базе аналитической платформы «Форсайт»

300+
каналов

40+ млн
зрителей

50+
сервисов
оператора

20+
характеристик
зрителей

Опыт использования

Проект «Медиаизмерения»

Результат

- Воспроизведен функционал старого ПО по расчётам панельных данных
- Разработан и внедрён функционал по обработке Big Data
- Подключены новые источники данных
- Настроен доступ в личные кабинеты
- Реализована визуализация отчётов
- Увеличена точность измерений нишевых тематических каналов



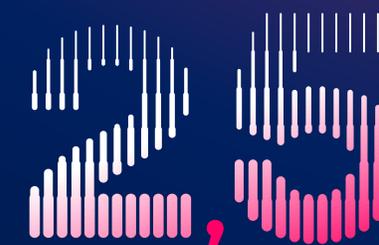
**млрд записей
в БД**



OLAP кубов

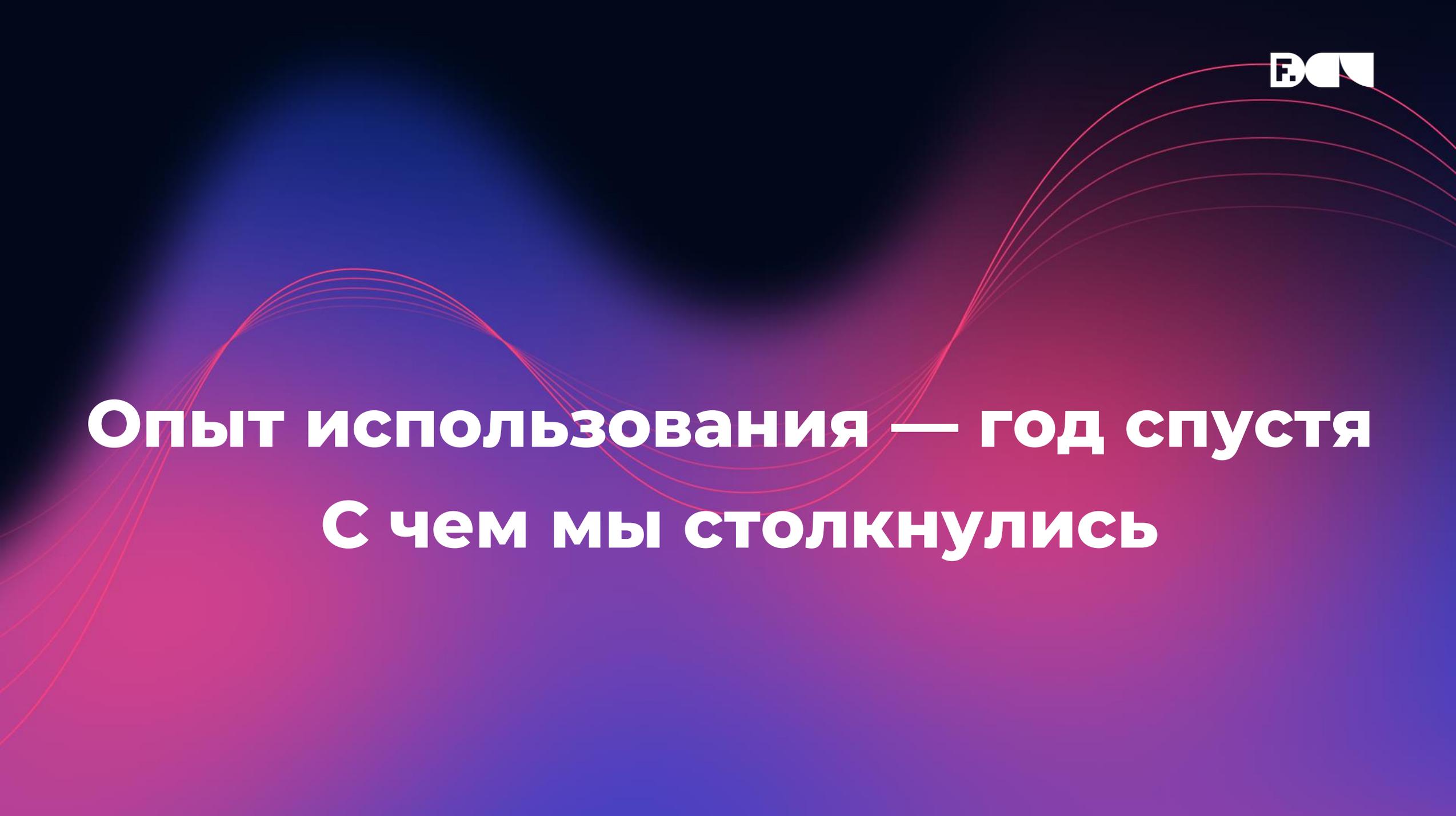


**параметров
на куб**



**млрд записей
на куб**



The background features a gradient from dark blue at the top to purple at the bottom. Several thin, white, wavy lines curve across the page, starting from the left and ending on the right, creating a sense of motion and flow.

Опыт использования — год спустя

С чем мы столкнулись

НОВЫЙ ВЫЗОВ

Низкая скорость построения отчётов при обработке Big Data

Рост объёма данных:



**раз больше
зрителей**



**объём данных
за сутки**



**обрабатываемые
параметры**

> 20 000 000 000
СОБЫТИЙ



НОВЫЙ ВЫЗОВ

Низкая скорость построения отчётов при обработке Big Data

Особенности работы выбранной в 2017 году СУБД — PostgreSQL

Охватные статистики

Для построения отчётов с охватными метриками необходимо посчитать всех уникальных зрителей (свыше 1 000 000 зрителей)

Нельзя оптимизировать

Этот объём данных с помощью текущих ПО и СУБД невозможно сжать, предрассчитать (в отличие от других метрик)

Работает с данными построчно

- Универсальная
- Проверенная
- Надёжная
- Подходит для обработки больших данных
- Нет доступной альтернативы

Как оптимизировать расчёты?

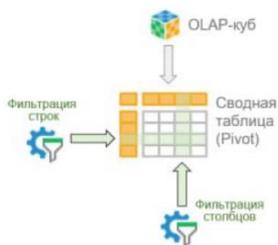
Виды фильтрации

Почему декарт может быть губителен для сводной таблицы и как с этим бороться?



ВИДИМЫЕ ДАННЫЕ СВОДНОЙ ТАБЛИЦЫ (наличие данных)

Фильтрация происходит на основе проверки «плоских» данных, которые были сформированы в сводной таблице (Pivot) из многомерного OLAP-куба

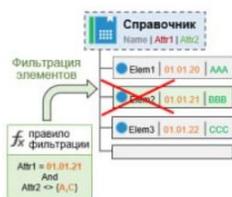


ЗНАЧЕНИЯ АТРИБУТОВ В ОДНОМ ИЛИ НЕСКОЛЬКИХ ИЗМЕРЕНИЯХ (устойчивые связи)

ОДНО ИЗМЕРЕНИЕ

по условию

В сводную таблицу не попадают элементы измерения, которые не соответствуют условию



НЕСКОЛЬКО ИЗМЕРЕНИЙ

по связям атрибутов

В сводную таблицу попадают только те элементы, которую удовлетворяют правилу взаимного равенства связей между значениями их атрибутов



по взаимной отметке для измерений

Для каждого элемента одного измерения заранее определяется набор отметок (SetSet), детализирующих состав элементов из других измерений



ВНЕШНИЕ ДАННЫЕ (динамически меняющиеся связи)

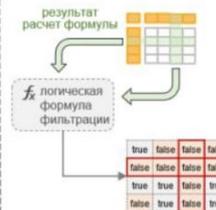
Фильтр-куб

Фильтрация строк боковика осуществляется на основе данных во внешнем кубе (фильтр-куб)



Расчетная формула / функция

Фильтрация строк или столбцов выполняется по заданной произвольной формуле, заданной пользователем



Решение — Переход на СУБД ClickHouse

Было

Postgres PRO

Метрика

Количество на PANEL

RPD

Количество фактовых таблиц куба на запросе

7

7

Количество строк в основной витрине

Почасовой отчёт:

50 млн строк + 50 тыс. в сутки

Почасовой отчёт:

5,2 млрд строк +11 млн в сутки

Поминутный отчёт:

145 млн строк + 700 тыс. в сутки

Поминутный отчёт:

6,2 млрд строк + 237 млн в сутки

Количество измерений

~ 25

~ 25

Количество параметров

~ 30

~ 30

Скорость выполнения типового расчёта

Почасовой отчёт:

~ 10 – 15 сек.

Почасовой отчёт:

~ 1,5 – 2.5 часа

Поминутный отчёт:

~ 30 – 60 сек.

Поминутный отчёт:

не считался, системы зависала

Стало

PostgreSQL (foreign tables on Clickhouse)

Метрика

Количество на PANEL

RPD

Количество фактовых таблиц куба на запросе

7

7

Количество строк в основной витрине

Почасовой отчёт:

50 млн строк + 50 тыс. в сутки

Почасовой отчёт:

5,2 млрд строк + 11 млн в сутки

Поминутный отчёт:

145 млн строк + 700 тыс. в сутки

Поминутный отчёт:

6,2 млрд строк + 237 млн в сутки

Количество измерений

~ 25

~ 25

Количество параметров

~ 30

~ 70

Скорость выполнения типового расчёта

Почасовой отчёт:

~ 10 – 15 сек.

Поминутный отчёт:

~ 30 – 60 сек.

Почасовой отчёт:

~ 10 – 25 сек.

Поминутный отчёт:

~ 10 – 45 сек.

Решение сработало

Скорость выполнения типового расчёта
на приёмниках с обратной связью

Предварительные замеры

x80 – x100 раз

Фактический прирост

5000



Новые задачи для системы

Развитие проекта

Deep analytics

Анализ аудитории по различным критериям

Портрет абонента

Аналитика использования сервисов

Поведение зрителей, их предпочтения

Жизненный цикл абонента, предсказание оттока

Рекомендательная система





A-TRAFFIC

Programmatic решение Агентства 2

Таргетированные и догоняющие рекламные кампании аудитории Триколора по такой же и похожей аудитории Рунета

Триколор

ID абонента
Биллинговые данные

Гибридная панель A2

Знания об абоненте



A-Metrics

Формирование целевой аудитории под запрос

Сотни параметров

Базовые сегменты:

доход, пол, возраст, семейное положение, состав домохозяйства

Расширенные сегменты:

Интересы, наличие домашних животных, спортивные предпочтения

DMP A-Traffic

Формирование аналогичной целевой аудитории во внешних средах

Внешние DMP

Поставщики данных

1000+ сегментов

Cookies

SSP A-Traffic

Кросс – платформенные рекламные кампании



Развитие проекта

Таргетирование рекламы по модели DSP/SSP/DMP

Выбранные фильтры

- Населенный пункт (город 100+)
- Тип устройства (SmartTV)
- ID устройства
- Пол (Мужской)
- Возраст (18-25 лет)
- Образование (среднее, высшее)
- Занятость (студент, работает)
- Доход (средний)
- Наличие детей (нет)

Паттерн потребления

- Временные интервалы (пн-чт 20-22, вс 22 -24)
- Программы (Фантастика, путешествия, развлечения)

Рекламодаватель

**Целевая аудитория
(вариант настройки
фильтров)**

**Определение
паттернов
потребления**

**Индивидуальный медиаплан
для каждого ID
удовлетворяющего фильтрам ЦА**

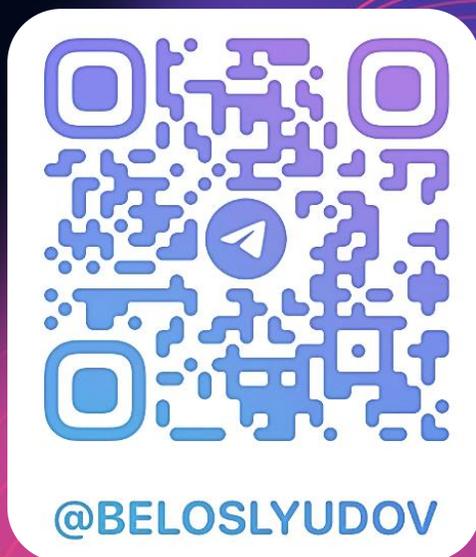
Выбранные фильтры

- Населенный пункт (город 400+)
- Тип устройства (IOS)
- ID устройства
- Пол (Женский)
- Возраст (15-30 лет)
- Образование (среднее, высшее)
- Занятость (учится, работает)
- Доход (средний)
- Наличие детей (нет)

Паттерн потребления

- Временные интервалы (пн-ср 19-20)
- Программы (Мелодрамы, путешествия, развлечения, музыка)

Есть вопросы?



Денис Белослюдов

Директор Департамента
специальных проектов
«Агентство 2»

+7 911 773 – 35 - 14

d.beloslyudov@agency2.ru

