avito.tech

Москва — 2021

DWH как продукт

Евгений Николаев

Руководитель юнита DWH

Евгений Николаев

Руководитель юнита DWH

- выпускник ВМК МГУ
- 6+ лет опыта разработки
- 3+ года опыта руководства
- фанат классных продуктов
- капитан ФК Авито
- КМС по шахматам
- говорю по-испански

Авито

40,8м

месячная аудитория

Больше четверти населения России

77м

объявлений

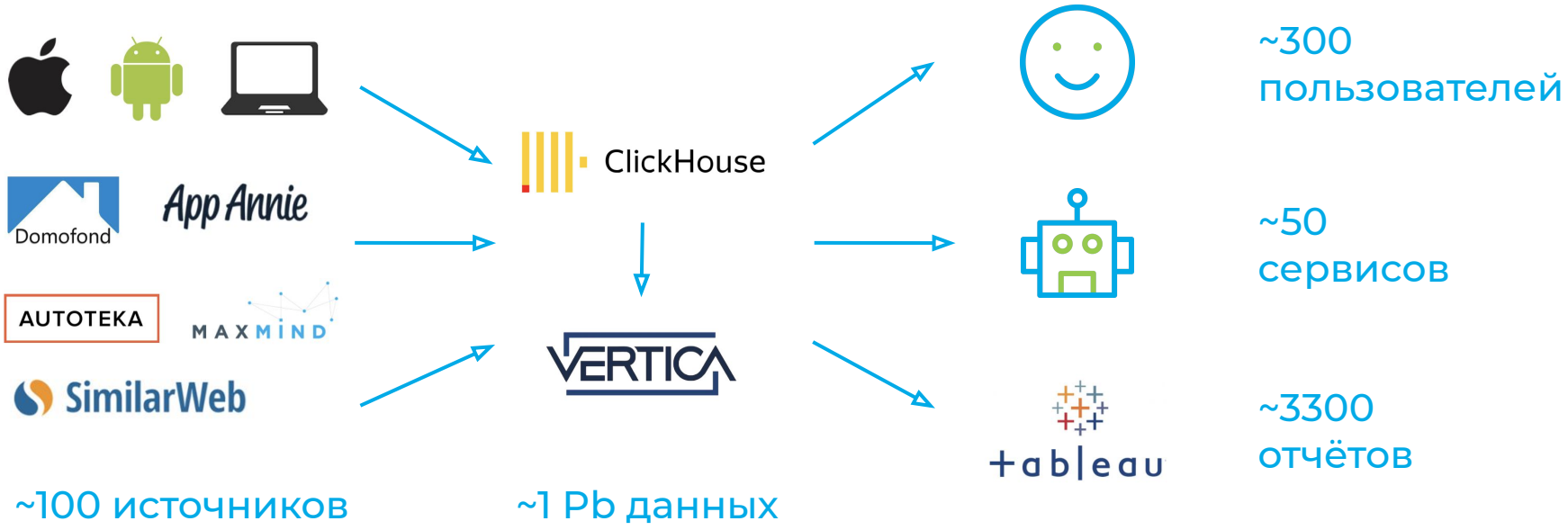
Каждый день создаётся почти 600К новых объявлений

182м

сделок в 2020

Это 60% от всех сделок в России

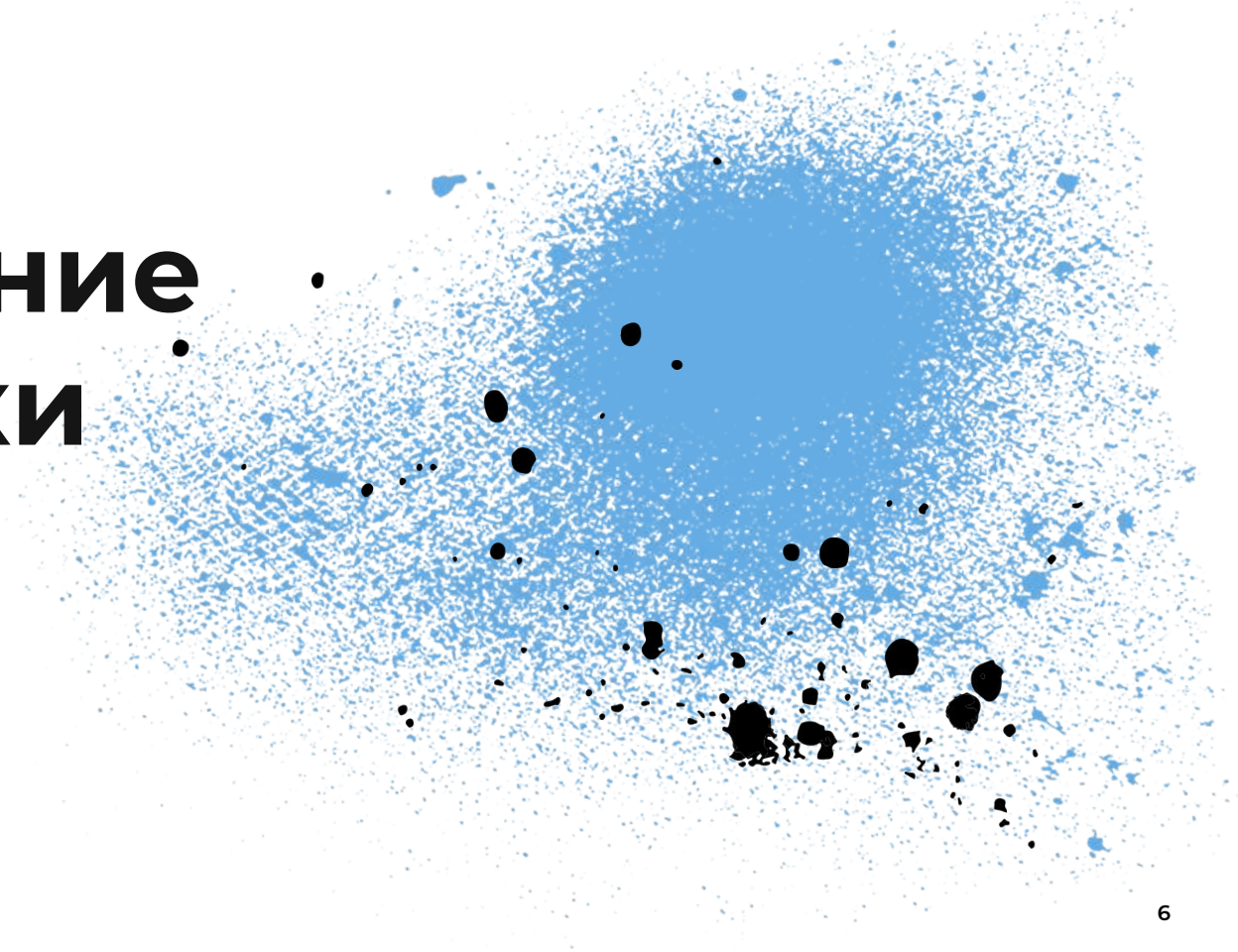
DWH в Авито


Domofond
App Annie
AUTOTEKA
MAXMIND
SimilarWeb
ClickHouse
VERTICA
+ableau
~300 пользователей
~50 сервисов
~3300 отчётов
~100 источников
~1 Pb данных

План выступления

- Зарождение аналитики в Авито
- Платформизация
- DWH как продукт
- Масштабирование
- Выводы и планы

Зарождение аналитики в Авито



Предпосылки

- Монетизация данных.
- A/B-тестирование.
- Автоматическая модерация.
- Финансовая отчётность.

  ⇨ **расширяемая BI-платформа.**

С чего мы начинали?

- Vertica + Tableau.
- ~50 Tb данных.
- ~6 дата-инженеров.
- ~25 пользователей-аналитиков.

кейс Vertica в Авито

Пример задачи: витрины

Analytical Processing / OLAP-2904

Подготовить расширенную витрину по активным пользователям

Description

К текущей витрине current_user добавить следующие поля:

- Item_Location_id самый частый локейшн среди 5 последних размещенных айтемов
- Item_Category_id самая частая категория среди 5 последних размещенных айтемов
- Item_Active - количество активных объявлений у юзера (External_id статус<10)
- Item_Total - общее количество объявлений у юзера
- Item_Blocked_Total - общее количество заблокированных объявлений у пользователя
- Item_Blocked_1W - количество заблокированных объявлений за последние 7 дней
- Item_New_1M - количество новых объявлений за последние 30 дней
- Item_New_1W - количество новых объявлений за последние 7 дней
- VAS_Amount_Total - объем потраченных средств на VAS за всю историю
- VAS_Amount_1M - объем потраченных средств на VAS за последний месяц
- VAS_Amount_1W - объем потраченных средств на VAS за последнюю неделю
- VAS_Count - количество VAS транзакций за всю историю

- Last_visit_time - датавремя последнего действия юзера на сайте

Пример задачи: ETL

Analytical Processing / OLAP-67

Необходимо создать свойство multi_vas для события 273 на step = 1

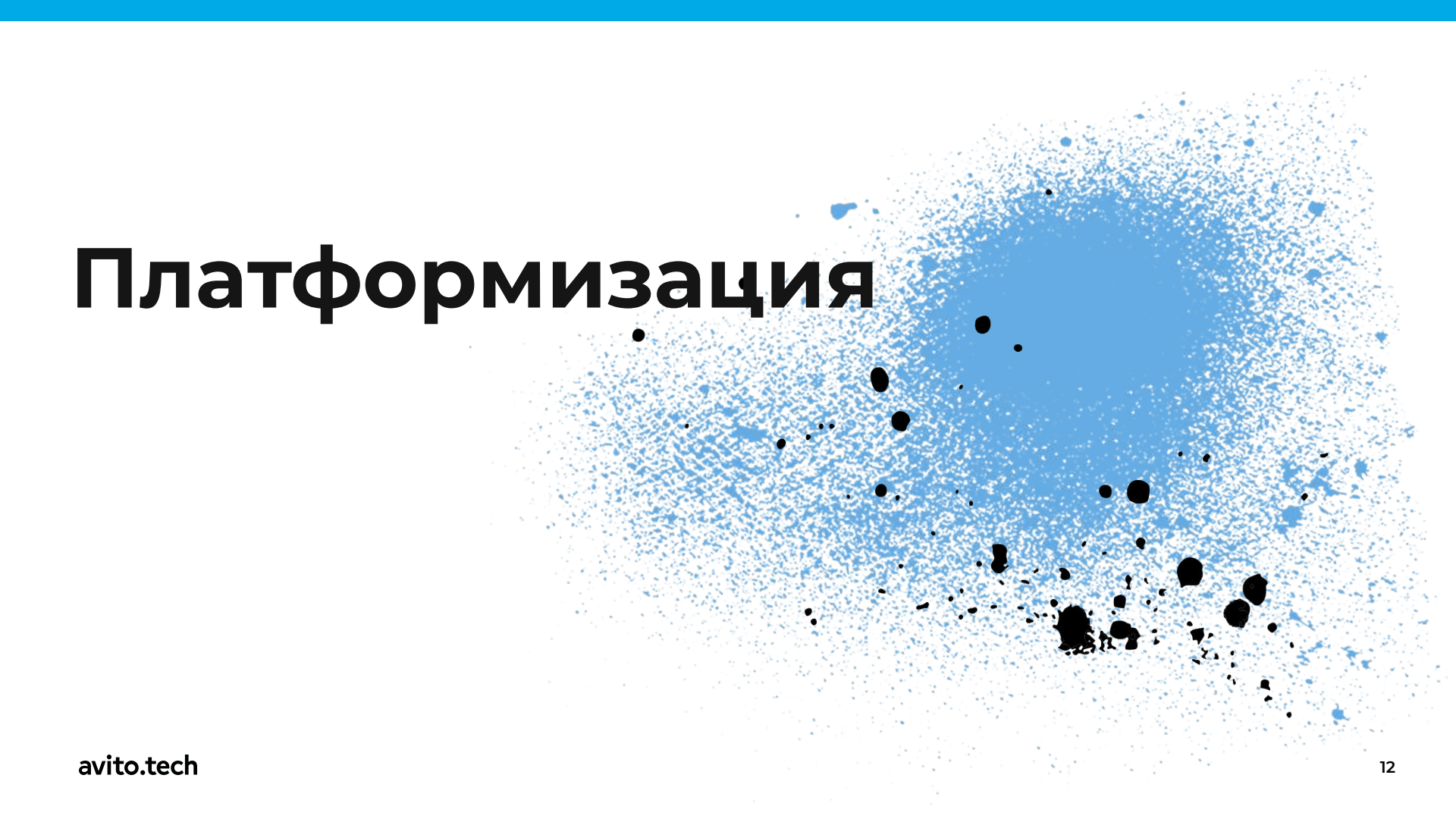
Description

Нужен сателит, который позволял бы понимать, что данное применение vas включает в себя несколько vasservice (dds.S_WebLog_VASService), то есть является мультиплатежом. Свойство нужно видеть на первом шаге воронки (dds.S_WebLog_Step), когда платежу еще не присвоен идентификатор трансакции (а значит нет идентификатора мультиплатежа - vasgroup), но по факту пользователь уже добавил в корзину несколько услуг.

Проблемы

- Как справиться с ростом задач?
- Как не погрязнуть в рутине?
- Как масштабировать команду?
- Как ставить цели?

Платформизация

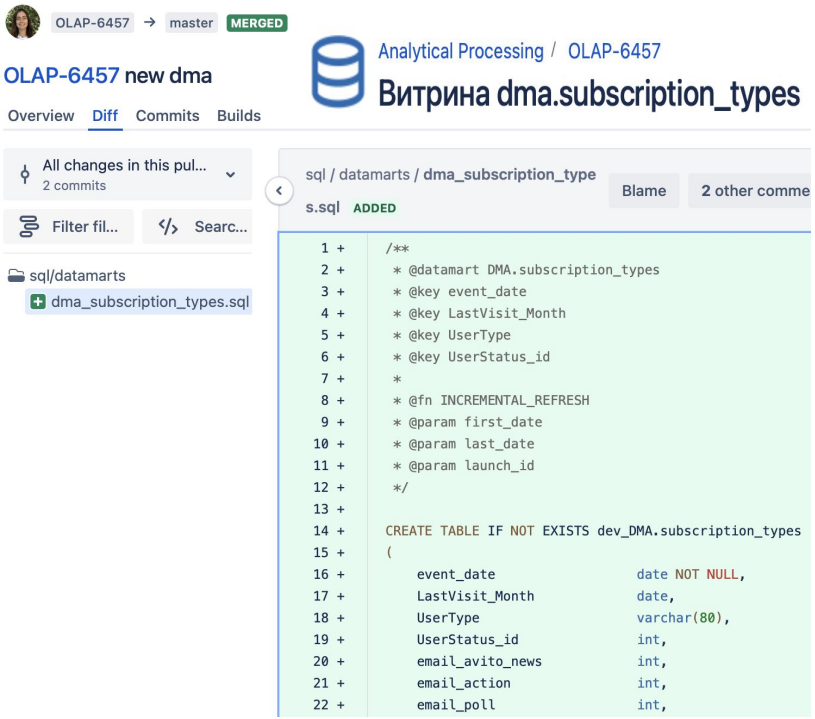


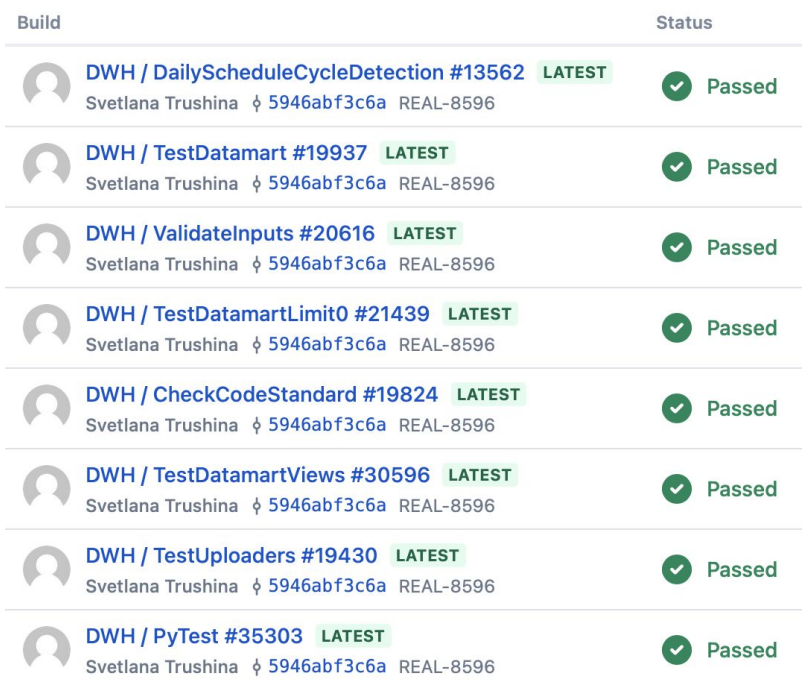
В чём идея платформизации?

- Мы создаём инструменты.
- Пользователи сами решают задачи.
- Мы помогаем им, где не получается.

Цель — 100% задач решаются пользователями.

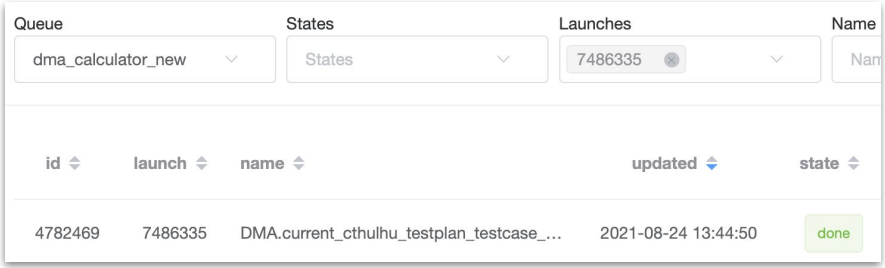
Витрины создают аналитики!





Datamart Framework

- SQL-based синтаксис.
- Инструмент оркестрации.
- Итог: 800+ ежедневных расчётов.




```
/**
 * @datamart DMA.subscription_types
 * @key event_date
 * @key LastVisit_Month
 * @key UserType
 * @key UserStatus_id
 *
 * @fn INCREMENTAL_REFRESH
 * @param first_date
 * @param last_date
 * @param launch_id
 */
```

ETL делают аналитики!

Analytical Processing / OLAP-11555

Разложить в DDS данные из realty-ownership

Description

Что за данные:

Наименование сервиса:

realty-ownership

Бизнес-смысл данных:

Мы в ближайшее время запускаем MVP верификации собственника в RE. Соответственно из сервиса забираем данные и статусы для аналитики продукта.

Здесь подробно про верификацию описано
- https://cf.avito.ru/pages/viewpage.action?pageId=175181829

Сценарии использования:

Данные в DWH нужны, чтобы настроить аналитику по продукту верификации собственника в RE.

PR с описанием раскладки тут:

http://stash.msk.avito.ru/projects/BI/repos/avito-dwh/pull-requests/21556/overview

ETL Framework

внешние данные

экстракт

раскладка

```
@set service_name = 'service-bivrost-api'
@set launch_id = 12345

create external table paas_service_libs
(
    libName varchar(256),
    language varchar(256),
    version varchar(256),
    complianceLevel varchar(256),
    complianceComment varchar(1024),
    supportedRange varchar(256),
    isDirect boolean
)
as
copy with
source CurlSource(url='http://kappa.k.avito.ru/service-atlas/getServiceLibsWithStatus', method='POST', data='{"serviceName": "' || :service
parser FlatJson(path='result.internal');


insert into sae.paas_service_libraries
select hash(:service_name, libName), :service_name, *, now(), :launch_id
from paas_service_libs;


CREATE TRANSTORE paas_services ON sae_paas_services.v_paas_service_libraries
(
    L_PaaSServiceLibrary_PaaSService        WITH service_library_id, service_name,
    S_PaaSServiceLibrary_Name               WITH service_library_id, lib_name,
    S_PaaSServiceLibrary_Lang               WITH service_library_id, lib_lang,
    S_PaaSServiceLibrary_Version            WITH service_library_id, lib_ver        ORDER BY download_time,
    S_PaaSServiceLibrary_ComplianceLevel    WITH service_library_id, lib_comp_lvl   ORDER BY download_time NULLABLE,
    S_PaaSServiceLibrary_ComplianceText     WITH service_library_id, lib_comp_text  ORDER BY download_time NULLABLE,
    S_PaaSServiceLibrary_CompliantVersion   WITH service_library_id, lib_sup_ver    ORDER BY download_time NULLABLE,
    S_PaaSServiceLibrary_IsDirect           WITH service_library_id, lib_is_direct  ORDER BY download_time,
);
```

- Все загрузки в SQL.
- 1 загрузка = 1 скрипт.
- Скрипты с версиями в Гите.

Масштабирование команды

dwh-growth

Цель — растить скорость решения аналитических задач:

поиск данных, построение отчетов, проверка гипотез, ETL.

dwh-infra

Цель — развитие инфраструктуры:

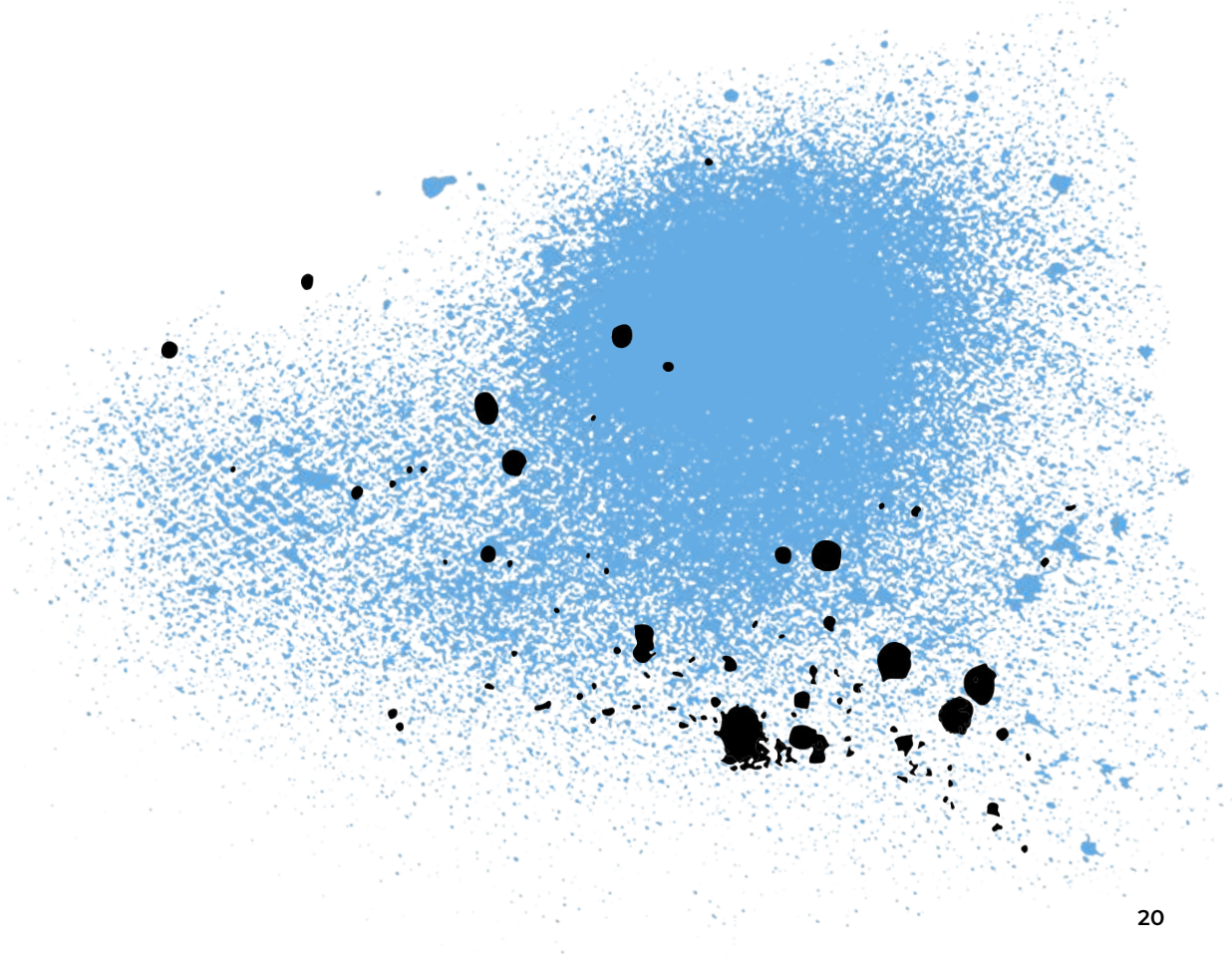
запросы, витрины, отчёты, антибот, realtime-аналитика.

Проблемы

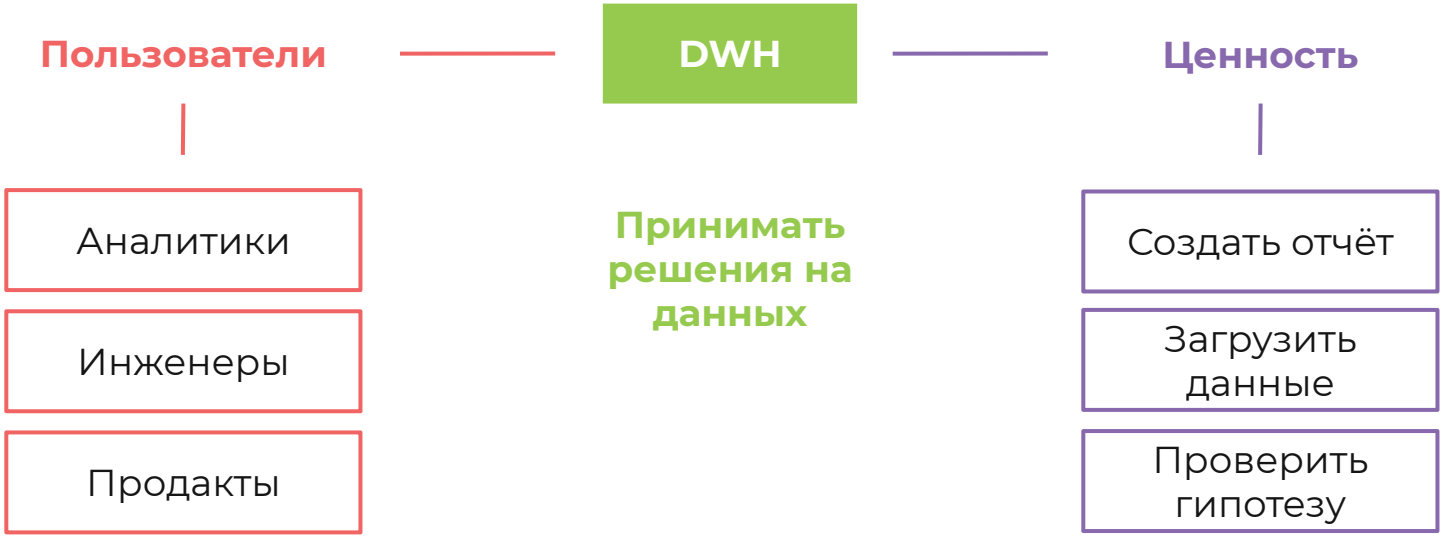
- dwh-growth: много пользовательских контекстов;
- dwh-infra: мало взаимодействия с пользователями;
- платформа медленно развивается;
- инфраструктура усложняется (+ClickHouse);
- больше бизнесовых пользователей;

  ⇨ юнит растёт.

DWH как продукт



DWH — продукт?


Пользователи
DWH
Ценность
Аналитики
Инженеры
Продакты
Принимать решения на данных
Создать отчёт
Загрузить данные
Проверить гипотезу

План

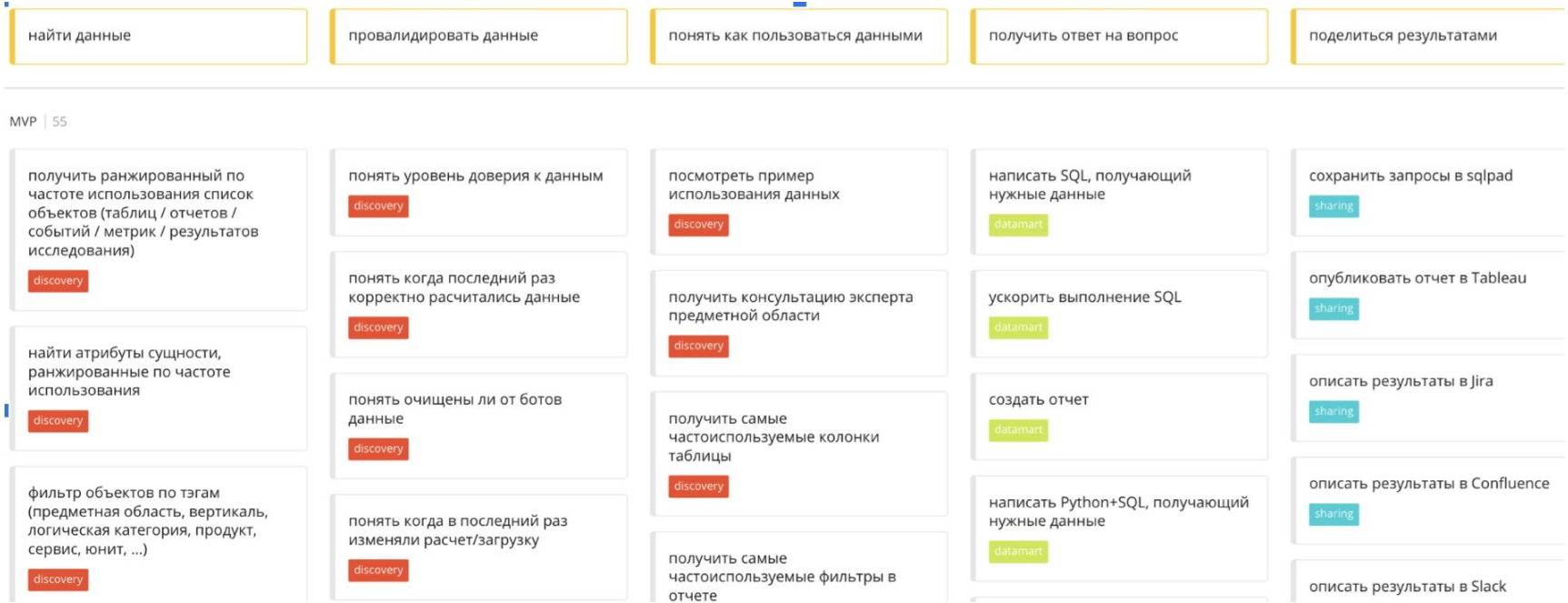
- Полезные практики:
  - User Story Mapping;
  - Customer Development.

- Примеры проблем:
  - поиск данных;
  - высокий порог входа в аналитику.

User Story Mapping

- Визуализировать и лучше представлять продукт.
- Что делаем?
- Для кого?
- Зачем?

Пример: проверка гипотезы



[плагин в miro]

Как составить?

- Определяем активности пользователя.
- Разбиваем активности на задачи.
- Нарезаем карту на куски по важности.
- Свадебный торт -> капкейки.

Jeff Patton лекция на ютубе

Jeff Patton «User Story Mapping» книга

Customer Development

- Как понять потребности пользователей?
- Как проверять гипотезы по развитию продукта?
- Как собирать обратную связь?

CustDev: примеры вопросов

- В чём для вас главная ценность продукта?
- 3 главных преимущества продукта.
- Что изменится, если продукта не будет?
- Какой продукт служит альтернативой?
- Вы рекомендовали знакомым, как?
- Есть ли факторы, которые мешают пользоваться?

CustDev: советы

- Все вопросы открытые.
- Правильно ли я понял, что «...»?
- Сценарий — канва разговора.
- Вопросы про прошлое, а не будущее.
- Интересуют факты, а не мнения и оценки.
- 5 почему.

книга Роберта Фитцпатрика

Проблема поиска данных

“Данные искать сложно. Чаще всего, приходится искать человека, который знает, где что лежит и как работает.”

“Часто приходится обращаться в чаты в слаке, чтобы понять, есть ли та или иная информация в вертике. Мне кажется, очень не хватает описания dds таблиц.”

Система поиска dwh-docs

agolovanev Aug 3rd at 12:11 PM
in #bi-and-dwh

всем привет, а где-то ведь наверняка хранятся кадастровые номера указанные в объявлениях? Кто , то может подсказать - где?)

2 replies

mbnekrasov 1 month ago

```
select * from tables where table_name
ilike '%cadastral%'
```

👍 1

agolovanev 1 month ago
спасибо, нашел

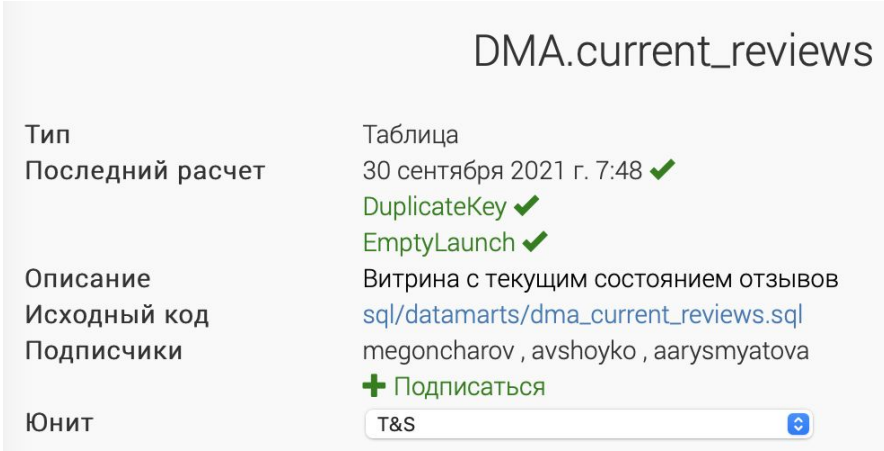
кадастровые номера

Искать везде | Все кластеры | Все схемы
Все юниты | Все типы таблиц | Все витрины

Название	Юнит	Описание
DDS.H_CadastralNumber	Common not Allocatable	Кадастровый номер
DDS.S_ItemCadastralNumberLink_IsDuplicate	-	Повторяется ли кадастрового номера на сайте в другом объявлении
DDS.S_ItemCadastralNumberLink_CheckBitMask	-	Битовая маска всех проверок, что происходили по кадастровому номеру
DDS.L_ItemCadastralNumberLink_CadastralNumber	-	Связка cadastralnumber с уникальной связкой item и кадастрового номера
DDS.L_ItemCadastralNumberLink_Item	-	Связка item с уникальной связкой item и кадастрового номера
DDS.S_CadastralNumber_VerifiedRosreestr	-	Подтверждение существование кадастрового номера от Росреестра
DDS.H_ItemCadastralNumberLink	Common not Allocatable	Ускуственный уникальный ключ между item и кадастровым номером
DDS.S_ItemCadastralNumberLink_IsShown	-	Будет ли показано на сайте, что кадастровый номер проверен

dwh-docs

- Поиск таблиц и отчётов.
- Документирование таблиц и колонок.
- Связи с Jira и Stash.
- Отображение аномалий в данных.
- Подписки на объекты.



Проблема DBeaver

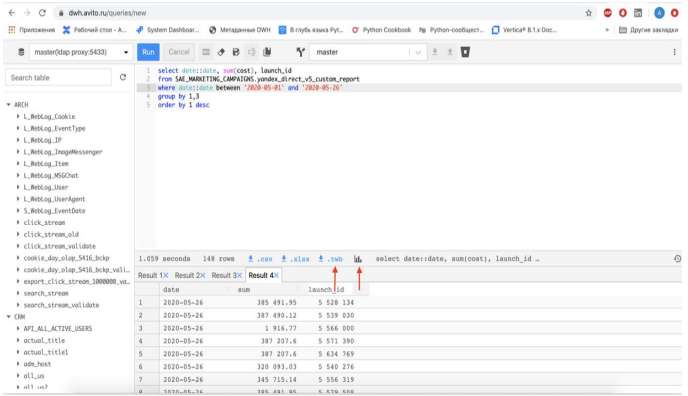
«Он стремный, убогий интерфейс, сложные настройки, после обновления периодически ломается (например, работа с параметрами), сложно было настраивать самому (аналитик помогал выбрать драйвер и ещё что-то), не так много документации/инфы в интернете/на stackoverflow.»

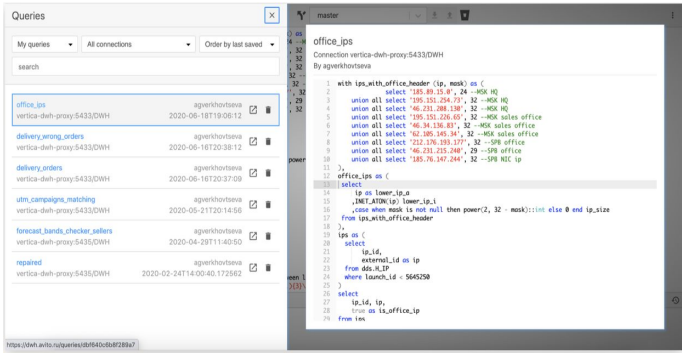
SQL-клиент в браузере

- Поддержка Vertica, CH, Postgres.
- Шаринг запросов через URL.
- Создание витрин.
- Работа с Tableau.
- Проверка планов запросов.

Обратите внимание на план запроса

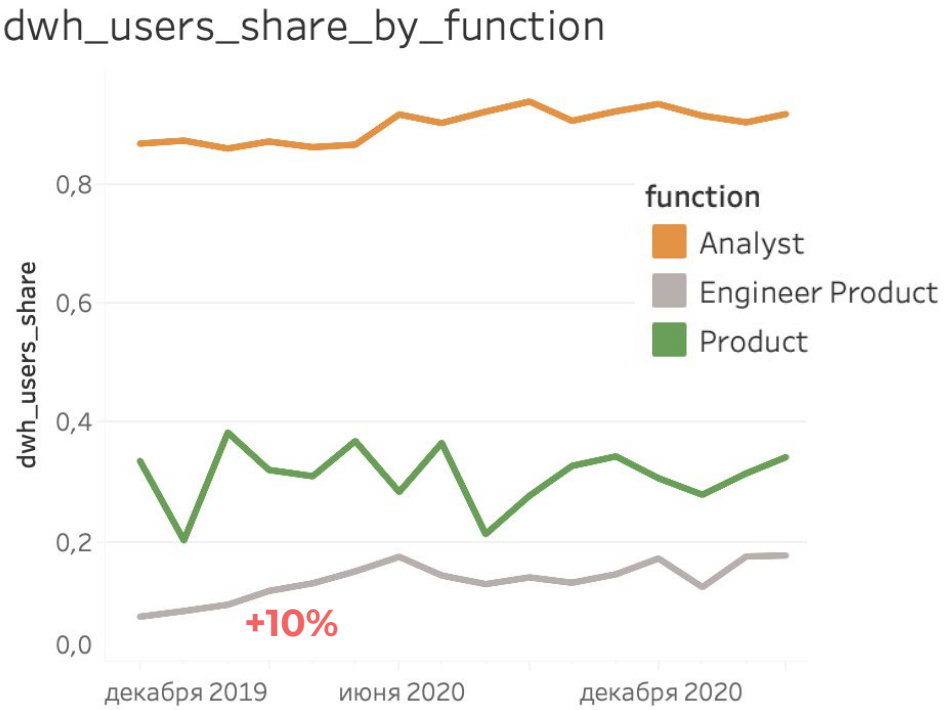
```
explain select count(*) from dma.current_item join dma.current_user

| +---> JOIN HASH [Cost: 509K, Rows: 2B] (PATH ID: 2) Outer (RESEGM
| | +-- Inner -> STORAGE ACCESS for current_user [Cost: 15K, Rows:
```





Итоги внедрения sqlpad

- Доля пользователей DWH среди инженеров выросла на 10%.
- Новые сотрудники используют только sqlpad.

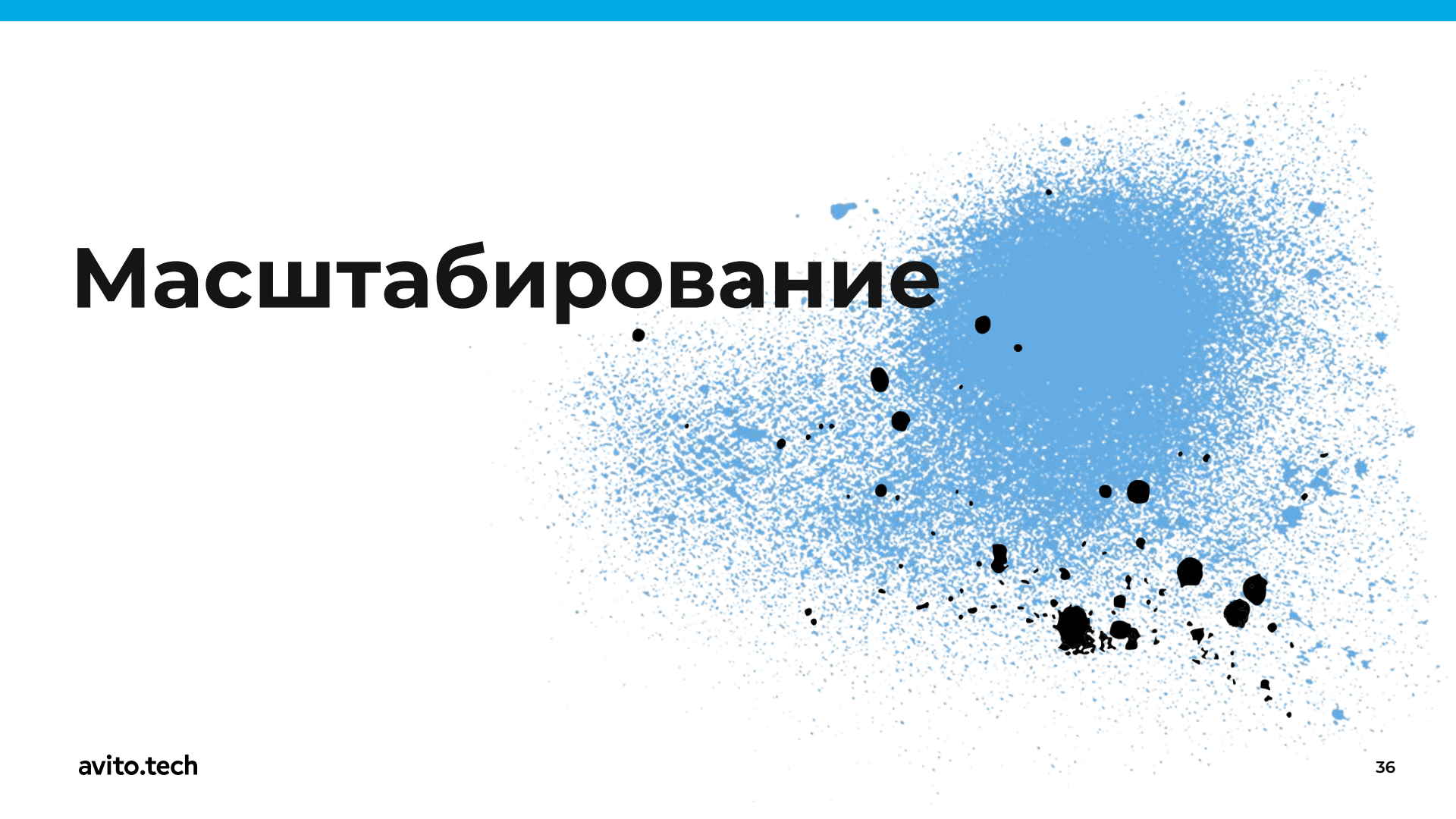


Полезные продуктовые практики

- User Story Mapping:
  - Jeff Patton лекция на ютубе
  - Jeff Patton «User Story Mapping» книга

- Customer Development:
  - книга Роберта Фитцпатрика

Масштабирование



Масштабирование

- Как сформировать цели команд?
- Как поделиться на команды?
- Как приоритизировать задачи?
- Как вовлечь бизнес-пользователей?

Цели команд

Activities из User Story Map
(высокоуровневые задачи пользователя):

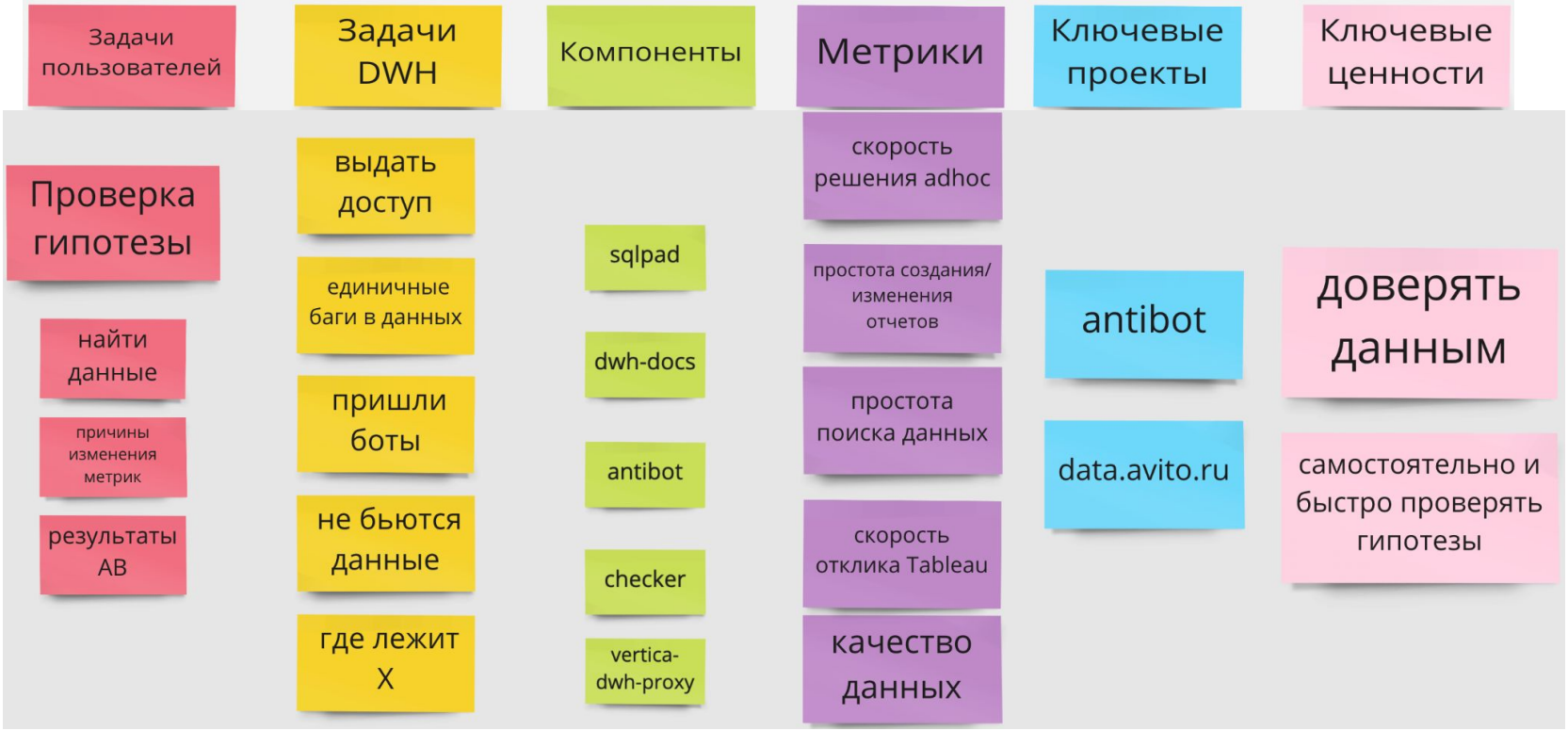
- Integration (интегрироваться с DWH).
- Datamart (создать регулярный отчёт).
- Usage (проверить гипотезу).

Integration

Datamart


Задачи пользователей
Задачи DWH
Компоненты
Метрики
Ключевые проекты
Ключевые ценности
Создание витрины
создать метрику
сделать рассылку
настроить выгрузку
новая витрина
правки витрины
пересчет витрины
метрика / разрез
klubok
DataMart Framework
uploader
репликация
tableau push
простота создания/ изменения витрин
скорость обновления данных
скорость пересчетов
простота написания запросов
скорость выполнения запросов
Datamart
metrics
recalc
самостоятельно и быстро изменять витрины
анализировать актуальные данные

Usage


Задачи пользователей
Задачи DWH
Компоненты
Метрики
Ключевые проекты
Ключевые ценности
Проверка гипотезы
найти данные
причины изменения метрик
результаты AB
выдать доступ
единичные баги в данных
пришли боты
не бьются данные
где лежит X
sqlpad
dwh-docs
antibot
checker
vertica-dwh-proxy
скорость решения adhoc
простота создания/ изменения отчетов
простота поиска данных
скорость отклика Tableau
качество данных
antibot
data.avito.ru
доверять данным
самостоятельно и быстро проверять гипотезы

Итоги разделения

- **Разнообразие задач**: рутина/вдохновляющая цель.
- **Фокус команд**: экспертиза, стимул автоматизации.
- **Прозрачность**: зоны ответственности.
- **Масштабирование**: пошарили взаимодействие с пользователями.

Как делиться?

- Определяем граничные условия (лиды, цели, число людей, роли).
- Предлагаем выбрать команду и роль самостоятельно.
- Делаем итеративно:
  - выбор команды;
  - оценка.
- Если оценки >= X — завершаем.

Советы

- Подробно объяснить процесс участникам и предпосылки. Ответственность — на команде.
- Кто-то не выберет команды: оставить время пообщаться и снять сомнения.
- Снять страх, что переход навсегда.
- Как вариант снятия ступора — предложить всем выписать приоритеты команд.

Как приоритизировать задачи?

Структура бэклога

- Квартальные OKR:
  - сбор кандидатов, оценки;
  - отбираем TOP, влезающий в 60% ёмкости.

- Спринтовые задачи:
  - 60%: OKR-задачи;
  - 20%: помощь аналитикам;
  - 20%: ретро AI + техдолг.

Как ранжировать задачи?

$$\frac{\text{Reach} \times \text{Impact} \times \text{Confidence}}{\text{Effort}}$$

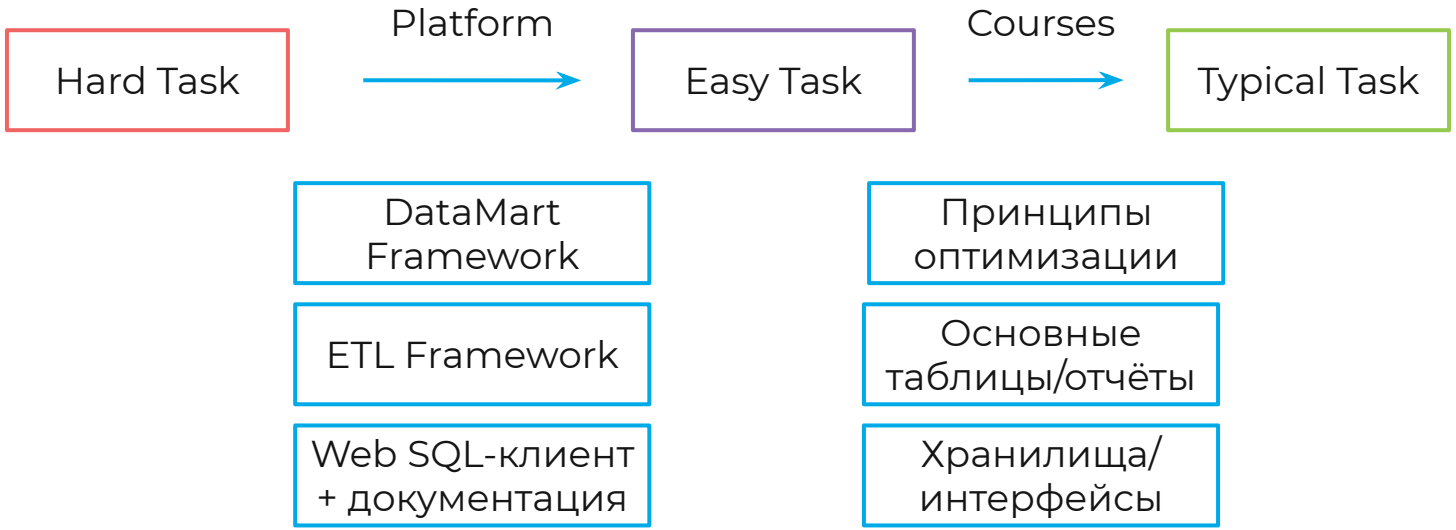
Использование RICE: советы

- Reach оценивать сложно.
- Холивар по ценности разных пользователей.
- Обсудить голосом перед голосованием.
- Откалибровать Impact/Effort (или отказаться?)

81 способ приоритизации фичей

Демократизация аналитики

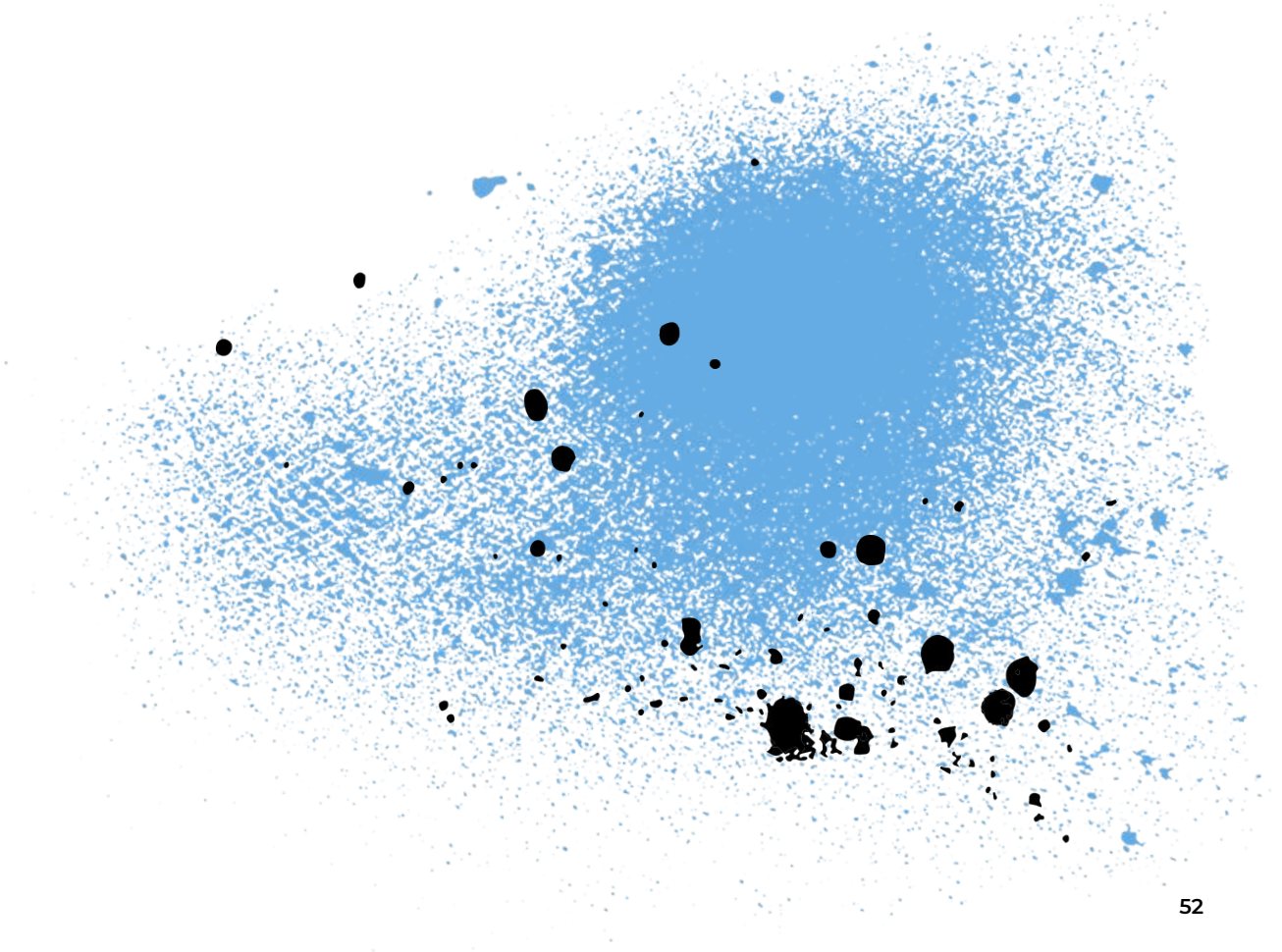
Упрощение задач



Масштабирование: выводы

- Как сформировать цели команд?
  - Activities из User Story Map.
- Как поделиться на команды?
  - Ребята сами решают.
- Как приоритизировать задачи?
  - RICE голосование.
- Как вовлечь бизнес-пользователей?
  - Упрощение задач, обучение.

Выводы



DWH как продукт

Производство данных

Аналитик-контрибьютор
Инженер

Загрузка данных
Построение отчётов

DWH

integration

datamart

usage

Использование данных

Продакт
Аналитик

Поиск данных
Проверка гипотез

Цифры

1 Pb

объём данных в CH & Vertica

Доступны для ежедневной аналитики

300

активных пользователей

Ежедневно решают задачи в хранилище данных

3300

отчетов в Tableau

Ежедневно обновляются по утрам

Планы

демократизация аналитики

Q4 2021

self-service: ETL, пересчет

доверие к данным

realtime-аналитика

Q1 2022

аналитика под ключ

opensource

DQ фреймворк

Q2 2022

IDE (SQL + Python + BI)

холодные данные

облачные СУБД

avito.tech

Москва — 2021

Евгений Николаев

Руководитель DWH

enikolaev@avito.ru

@nikolaevgenii

@eanikolaev

Если вы хотите построить корабль,
Не надо будоражить народ,
отправлять собирать древесину,
делить работу и отдавать приказы.
Лучше научите людей тосковать
По обширному и бесконечному морю.