RDDM Data Warehouse

All information in one place

О компании its.xyz

Компания its.xyz была основана в городе Санкт-Петербург, в 2015 году.

Группа инженеров – выпускников ведущих технических вузов со всей страны объединилась онлайн с целью совместной работы над наукоемкими задачами, возникающими в современном бизнесе.

Мы оптимизируем рутинные задачи и используем бережливое производство, с целью минимизации издержек на проверку гипотез. В работе над проектами мы комбинируем научный подход и современные управленческих практики, что позволяет достигать целей заказчика в кратчайшие сроки, одновременно сохраняя вовлеченность и мотивацию сотрудников.

ВкусВилл ВТБ РЖД Циан РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

В этих и многих других компаниях живет и работает код, написанный инженерами its.xyz

Экспертиза

DataScience, Machine Learning, Algorithm Development, Due Diligence, Architecture, Support, WEB Development, DevOps, Frontend

Технологии

Python, JavaScript, TypeScript, C++, C# OpenSearch, AWS, Linux, AstraLinux, Yandex.Cloud, Tensorflow, Keras, NextJS, ReactJS, VueJS, NuxtJS, Flask, FastAPI, TelegramAPI, Redis, Kafka, Celery, Redash, Postgres, Avro, Pandas и многие другие

Сайт и витрина проектов

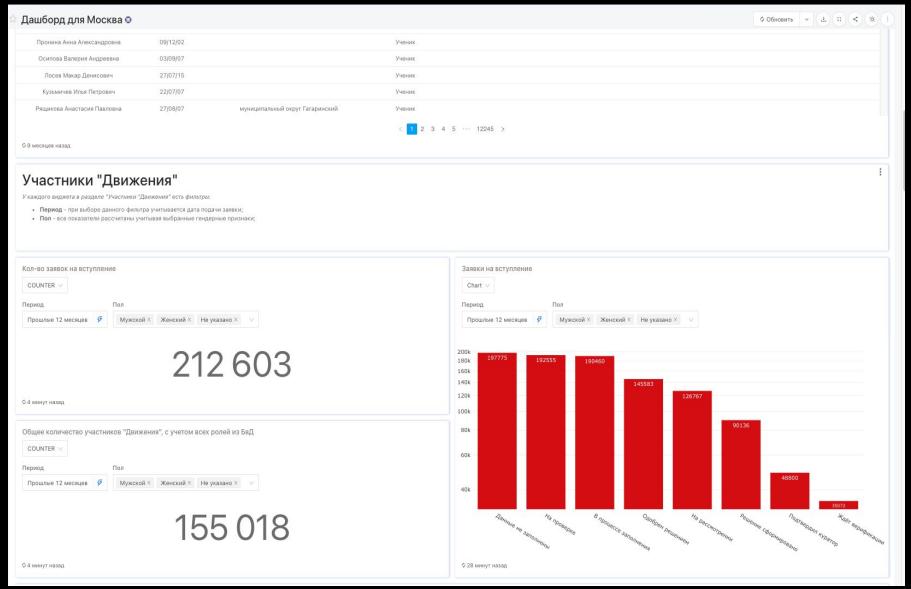
https://its.xyz

Описание бизнес задачи

Проблема
Отсутствие у заказчика инструмента для сбора, синхронизации и визуализации данных из разных сервисов экосистемы продуктов. Отчеты строятся вручную, нет портрета пользователя и автоматизации подсчета KPI. Отсутствуют отчеты, формируемые в реальном времени на актуальных данных, что крайне важно для принятия управленческих решений. Отсутствуют сводные аналитические отчеты для разных групп сотрудников организации, а также возможность выгружать большие объемы разнородных данных из разных источников.

Цель
Создать единое хранилище данных из разных информационных источников заказчика, унифицировать формат хранения, предоставить пространство для анализа данных с использованиям различных источников системы, предоставить удобный инструмент для визуализации данных, создать интерактивный инструмент построения отчетов с персонализированными настройками среды для каждого пользователя, добавить возможность использовать собственные ресурсы данных, которые не принадлежат общему хранилищу системы.

Точка внедрения
Разработка и внедрение ETL-пайплайнов для сбора, нормализации, хранения данных из экосистемы продуктов заказчика, а также внедрение инструмента для визуализации накопленных данных. Доработанное новыми функциями и возможностями open-source решение Redash, кастомизированное под нужды Заказчика.



Ценность для бизнеса и экспертиза

Область применения:
Корпоративная обработка больших данных. Финансовая и ресурсная оптимизация. Автоматизация построения регулярных отчетов, получение значений ключевых метрик в реальном времени и отслеживание портрета пользователя, объединяя данные между разными подсистемами

Ценность для бизнеса:
Стандарты качества в обработке данных – полная прозрачность в отслеживании ключевых показателей. Трейсинг дубликатов. Понятные и легкие инсайды. Устранение неоптимальностей. Обнаружение нарушителей. Точный KYC.

Экспертиза:
Инженеры компании its.xyz разработают архитектуру, методологию и последовательность сервисов ETL, хранения и визуализации данных. Бизнес аналитики нашей компании помогут найти первые инсайды и обучат вашу команду пользоваться данными и их анализировать.

Разработка и наполнение озера данных

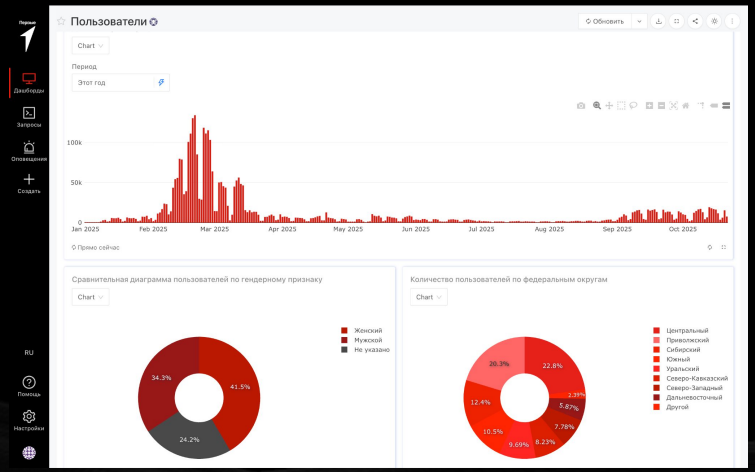
Проблематика: есть девять источников данных с уникальными пользователями. В них хранится информация о пользователях - участие в мероприятиях, привязка к региону, школе, фед. округу и т.д.. У каждого сервиса есть свои уникальные идентификаторы, не связанные друг с другом. А также имеется набор пересекающихся изменяемых полей.

Цель проекта: разработать механизм объединения данных из нескольких источников, нормализовать объединенные данные и визуализировать данные об установленных связях.

Метод достижения цели:
1. Разработать единую модель данных
2. Реализовать модуль хранения сырых данных из нескольких источников
3. Реализовать модуль нормализации данных
4. Реализовать модуль синхронизации данных
5. Реализовали модуль хранения нормализованных данных
6. Интегрировать сервис визуализации данных

Команда проекта (its.xyz):
1 – Проектный менеджер
2 - Инженера разработчика
1 - Бизнес аналитик
1 - Архитектор данных
1 - Старший DevOps инженер
1 – Ведущий инженер разработчик
1 - Старший инженер разработчик

Что было сделано:
– Провели фазу исследования в течение 2 недель, в рамках которой предоставили PoC, подтвердили гипотезу о возможности объединения данных из разных источников заказчика, предоставив единую модель данных
– Реализовали модуль хранения сырых данных из разных источников
– Реализовали модуль нормализации данных
Реализовали модуль хранения нормализованных данных
– Реализовали модуль синхронизации данных
– Интегрировали сервис визуализации данных



Таймлайн проекта:

Исследование	Реализация	Тест / Интеграция	Доработки/Тех. поддержка
0.5 мес	5 мес	3 мес	2 года

Далее подробнее о проекте

Основные этапы работ

Проектирование

- ✅ Провели технический анализ источников и выполнили проектирование ETL-пайплайнов;
- ✅ Провели анализ данных из источников, выявили пересекающиеся и уникальные поля, разработали механизм определения уникальных пользователей двух систем;
- ✅ Разработали механизмы детектирования и обработки ситуаций, когда данные о связях пользователя с остальными сущностями поступили на обработку раньше данных о пользователе.

Реализация

- ✅ Реализовали модуль хранения сырых данных с механизмом фиксации ошибок нескольких родов: когда родительская сущность не обнаружена в базе данных или запись из источника содержит неполную информацию для записи в нормализованный слой;
- ✅ Реализовали модуль нормализации данных:
  - Валидация данных на основе разработанной обобщенной схемы данных;
  - Механизм выявления и корректного обновления нескольких записей об одном и том же пользователе согласно выработанному уникальному ключу;
  - Механизм устранения технических ошибок в данных из источников.
- ✅ Реализовали модуль синхронизации данных с помощью брокера сообщений;
- ✅ Интегрировали open-source сервис визуализации данных,

Методология разработки

Agile / Scrum

Команда its.xyz при работе как на долгосрочных, так и краткосрочных проектах использует для работы гибкие методологии на базе фреймворка Scrum. В рамках данного проекта длительность спринта составила 1 неделю. Проект был укомплектован в 8 спринтов, после завершения и демонстрации результатов, заказчик принял решение о продлении еще на 8 спринтов. В рамках проекта проводились регулярные «дейли встречи» и «груминг». Команда принимала участие в наполнении беклога, приоритизации задач и генерации идей для конечного заказчика

Команда

its.xyz

1 – Проектный менеджер
2 – Инженер разработчик
1 – Бизнес аналитик
1 – Архитектор данных
1 – Старший DevOps инженер
1 – Ведущий инженер разработчик
1– Старший инженер разработчик

Со стороны заказчика

1 – Проектный менеджер
1 – Владелец продукта

Лог работ

Спринты (длительность 1 неделя)

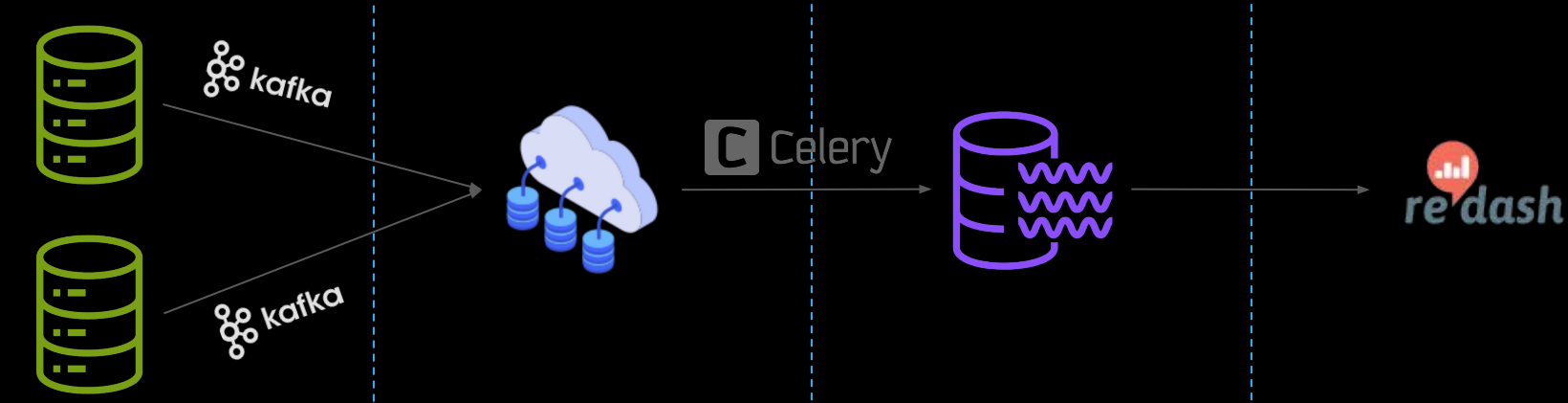
(1-2) Инициализация; разработка модуля хранения сырых данных

(3-4) Подключение к источникам данных; разработка схемы нормализованной схемы данных, построение ETL-процессов

(5-6) Наполнение озера данными; нормализация данных; построение дашбордов

(7-8) Тестирование и интеграция; приемо-сдаточные испытания

Технические особенности решения



Стек технологий

Управление проектом

Google Sheets GitLab Jenkins

Инфраструктура

kafka Linux docker Apache Airflow

redash Windows 10 Sentry MINIO

Программная среда

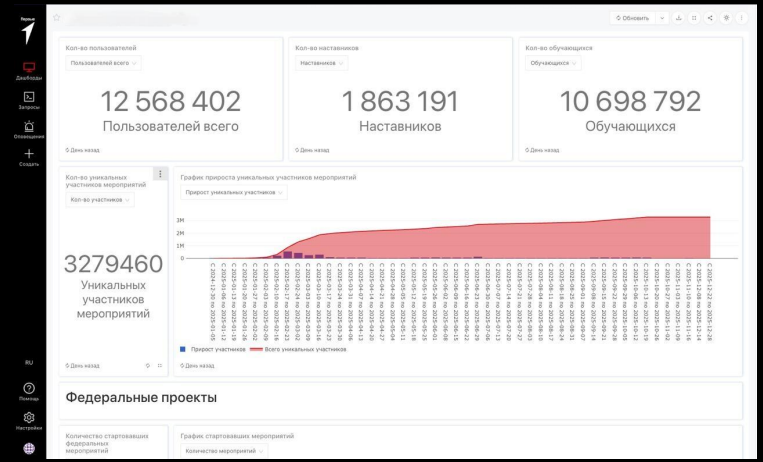
python PyCharm VS Code

Библиотеки

FastAPI Celery SQLAlchemy pandas

Flask

Ключевые результаты



Было

👎 несколько разных платформ с данными, каждая со своим уникальным набором данных, нет возможность собирать портрет пользователя из разных систем

Стало

✅ единая база данных, сформированная объединением данных из нескольких источников по уникальным пересекающимся полям

Метрика*	**До**	**После**
Количество источников данных	9	1
Количество записей о пользователях	~ 51 млн, с дубликатами	~ 13 млн, уникальных
Возможность визуализации данных в едином сервисе	👎	✅
Контроль над системой	Внешний	Внутренний

Было

👎 несколько инструментов визуализации данных для каждого источника

Стало

✅ единый сервис визуализации данных с полным контролем доступа и управления

Итоги

Провели технический анализ источников и данных из источников, спроектировали ETL-пайплайны

Разработали механизм детекции ошибочных данных и их прогрузки в случае устранения ошибок

Разработали механизмы нормализации сырых данных от заказчика

Разработали и реализовали механизмы синхронизации данных

Интегрировали сервис визуализации данных с полным контролем доступа

Контакты

Заказать сервис или разработку
presales@its.xyz

Стать партнером
partners@its.xyz

Сайт компании
its.xyz

Спасибо!