RDDM Data Warehouse

All information in one place

About its.xyz company

its.xyz company was founded in St. Petersburg, in 2015.

A group of engineers, graduates of leading technical universities from all over the country, joined together online to work together on high–tech tasks that arise in modern business.

We optimize routine tasks and use lean manufacturing to minimize the cost of hypothesis testing. When working on projects, we combine a scientific approach and modern management practices, which allows us to achieve the customer's goals in the shortest possible time, while maintaining employee engagement and motivation.

ВкусВилл ВТБ РЖД циан РОССИЯ – СТРАНА ВОЗМОЖНОСТЕЙ

The code written by its.xyz engineers lives and works in these and many other companies.

Expertise

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

Technologies

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

Website and showcase of projects

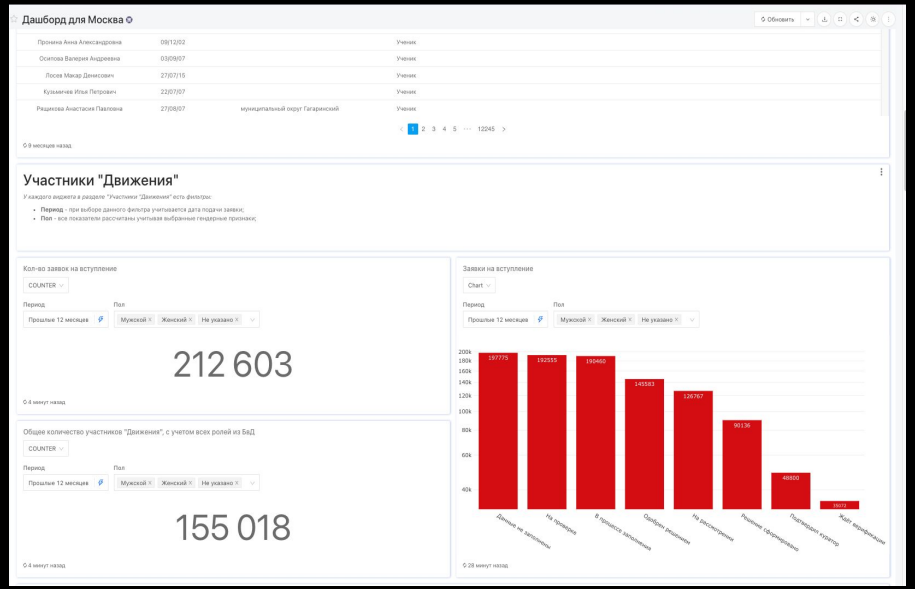
https://its.xyz

Business task description

Issue
The customer does not have a tool for collecting, synchronizing, and visualizing data from various services in the product ecosystem. The reports are built manually, there is no portrait of the user and no automation of calculating KPIs. There are no reports generated in real time based on up-to-date data, which is extremely important for making managerial decisions. There are no consolidated analytical reports for different groups of employees in the organization, as well as the ability to download large amounts of heterogeneous data from different sources.

Goal
Create a single data warehouse from different information sources of the customer, unify the storage format, provide space for data analysis using various sources of the system, provide a convenient tool for data visualization, create an interactive reporting tool with personalized environment settings for each user, add the ability to use your own data resources that do not belong to the common storage system.

The point of implementation
Development and implementation of ETL pipelines for collecting, normalizing, and storing data from the customer's ecosystem of products, as well as the implementation of a tool for visualizing accumulated data. The Redash solution, modified with new functions and features of the open-source, customized to the needs of the Customer.

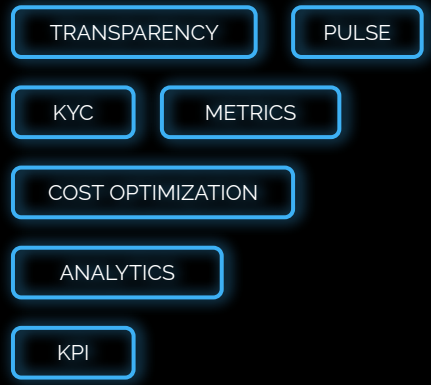


Business value and expertise

Scope of application:
Corporate big data processing. Financial and resource optimization. Automating the construction of regular reports, obtaining the values of key metrics in real time, and tracking the user's portrait by combining data between different subsystems.

Business value:
Quality standards in data processing – full transparency in tracking key indicators. Duplicate tracking. Clear and easy insights. Elimination of suboptimalities. Detection of intruders. The exact KYC.

Expertise:
its.xyz engineers will develop the architecture, methodology, and sequence of ETL, data storage, and visualization services. Our company's business analysts will help you find the first insights and train your team to use and analyze the data.



Development and filling of the data lake

Problematics: there are nine data sources with unique users. They store information about users, such as participation in events, location, school, and federation. Each service has its own unique identifiers that are not related to each other. And there is also a set of overlapping mutable fields.

The purpose of the project: to develop a mechanism for combining data from multiple sources, normalize the combined data, and visualize data on established relationships.

The method of achieving the goal:
1. Develop a unified data model
2. Implement a module for storing raw data from multiple sources
3. Implement the data normalization module
4. Implement a data synchronization module
5. Implement a normalized data storage module
6. Integrate the data visualization service

The project team (its.xyz):
1 – Project Manager
2 - Development Engineer
1 - Business Analyst
1 - Data Architect
1 - Senior DevOps Engineer
1 – Lead Development Engineer
1 - Senior Development Engineer

What was done:
– Conducted a research phase for 2 weeks, during which they provided a PoC, confirmed the hypothesis about the possibility of combining data from different customer sources by providing a single data model
– Implemented a module for storing raw data from different sources
– Implemented a data normalization module - Implemented a module for storing normalized data
– Implemented a data synchronization module
– Integrated a data visualization service

Project timeline:

Study	Realization	Test / Integration	Improvements/Technical support
0.5 months	5 months	3 months	2 years

Further details about the project

The main stages of the work

Designing

- ✅ We conducted a technical analysis of the sources and performed the design of ETL pipelines;
- ✅ We analyzed data from sources, identified overlapping and unique fields, and developed a mechanism for identifying unique users of the two systems;
- ✅ We developed mechanisms for detecting and processing situations when data about the user's relationships with other entities was received for processing before data about the user.

Realization

- ✅ We implemented a raw data storage module with a mechanism for fixing errors of several kinds: when the parent entity is not found in the database or the record from the source contains incomplete information for writing to the normalized layer;
- ✅ Implemented a data normalization module:
  - Data validation based on the developed generalized data schema;
  - A mechanism for identifying and correctly updating multiple records about the same user according to a developed unique key;
  - A mechanism for eliminating technical errors in data from sources.
- ✅ Implemented a data synchronization module using a message broker;
- ✅ Integrated an open-source data visualization service.

Development methodology

Agile / Scrum

When working on both long-term and short-term projects, the its.xyz team uses flexible methodologies based on the Scrum framework. Within the framework of this project, the duration of the sprint was 1 week. The project was completed in 8 sprints, after completion and demonstration of the results, the customer decided to extend it for another 8 sprints. Regular "daily meetings" and "grooming" were held as part of the project. The team participated in filling out the backlog, prioritizing tasks and generating ideas for the end customer.

Team

its.xyz

1 – Project Manager
2 – Development Engineer
1 – Business Analyst
1 – Data Architect
1 – Senior DevOps Engineer
1 – Lead Development Engineer
1– Senior Development Engineer

From the customer's side

1 – Project Manager
1 – The product owner

Work log

Sprints (duration 1 week)

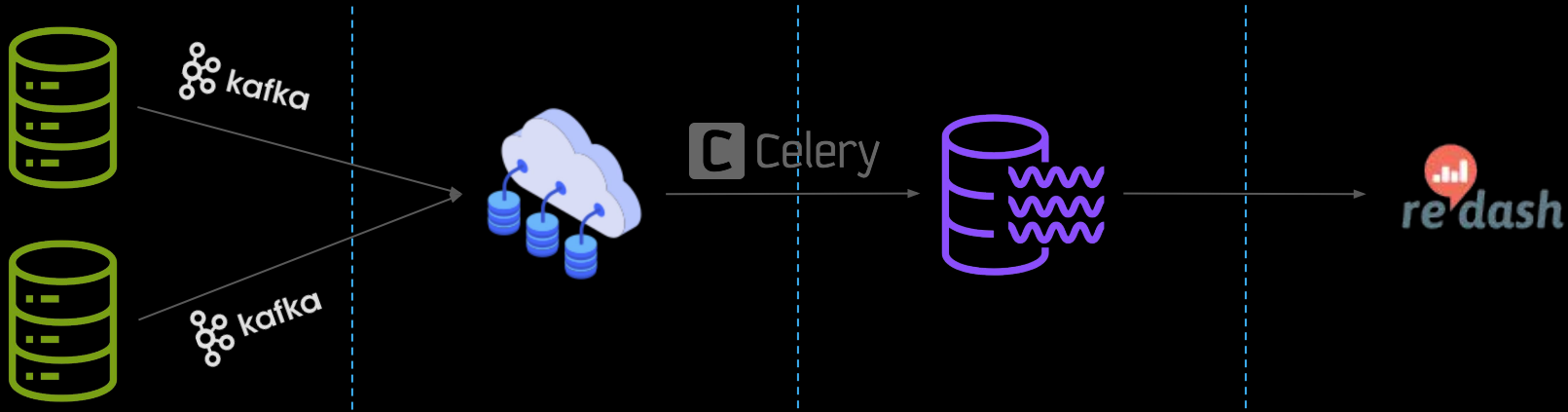
(1-2) Initialization; development of a raw data storage module

(3-4) Connection to data sources; development of a normalized data schema, building ETL processes

(5-6) Filling the lake with data; normalization of data; building dashboards

(7-8) Testing and integration; acceptance tests

Technical features of the solution



Technology stack

Project management

Google Sheets GitLab Jenkins

Infrastructure

kafka Linux docker Apache Airflow

redash Windows 10 Sentry MINIO

Software environment

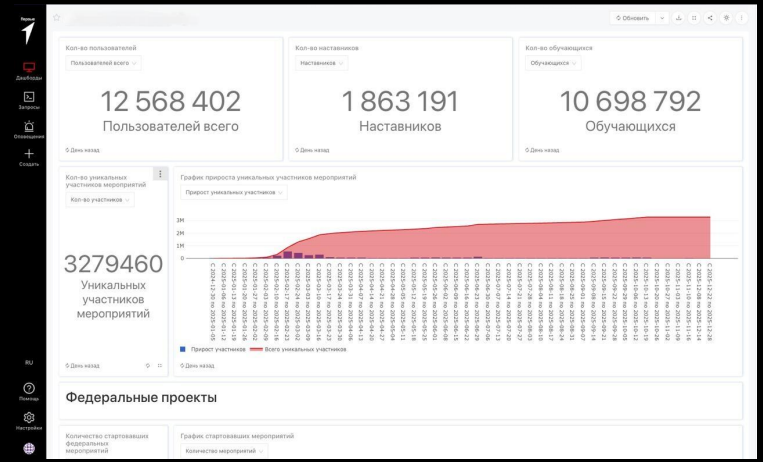
python PyCharm VS Code

Libraries

FastAPI Celery SQLAlchemy pandas

Flask

Key results



Metric*	Before	After
Number of data sources	9	1
Number of user records	~ 51 million, with duplicates	~ 13 million unique ones
The ability to visualize data in a single service	👎	✅
System control	External	Internal

It was

👎 There were several different data platforms, each with its own unique set of data, there was no way to collect a user's portrait from different systems

Become

✅ a single database formed by combining data from multiple sources on unique overlapping fields

It was

👎 multiple data visualization tools for each source

Become

✅ a single data visualization service with full access and management control

Results

We conducted a technical analysis of sources and data from sources, designed ETL pipelines

We developed a mechanism for detecting erroneous data and uploading it in case of error correction.

We developed mechanisms for normalizing raw data from the customer

We developed and implemented data synchronization mechanisms

Integrated a data visualization service with full access control

Contacts

Order a service or development
presales@its.xyz

Become a partner
partners@its.xyz

Company website
its.xyz

Thank you!