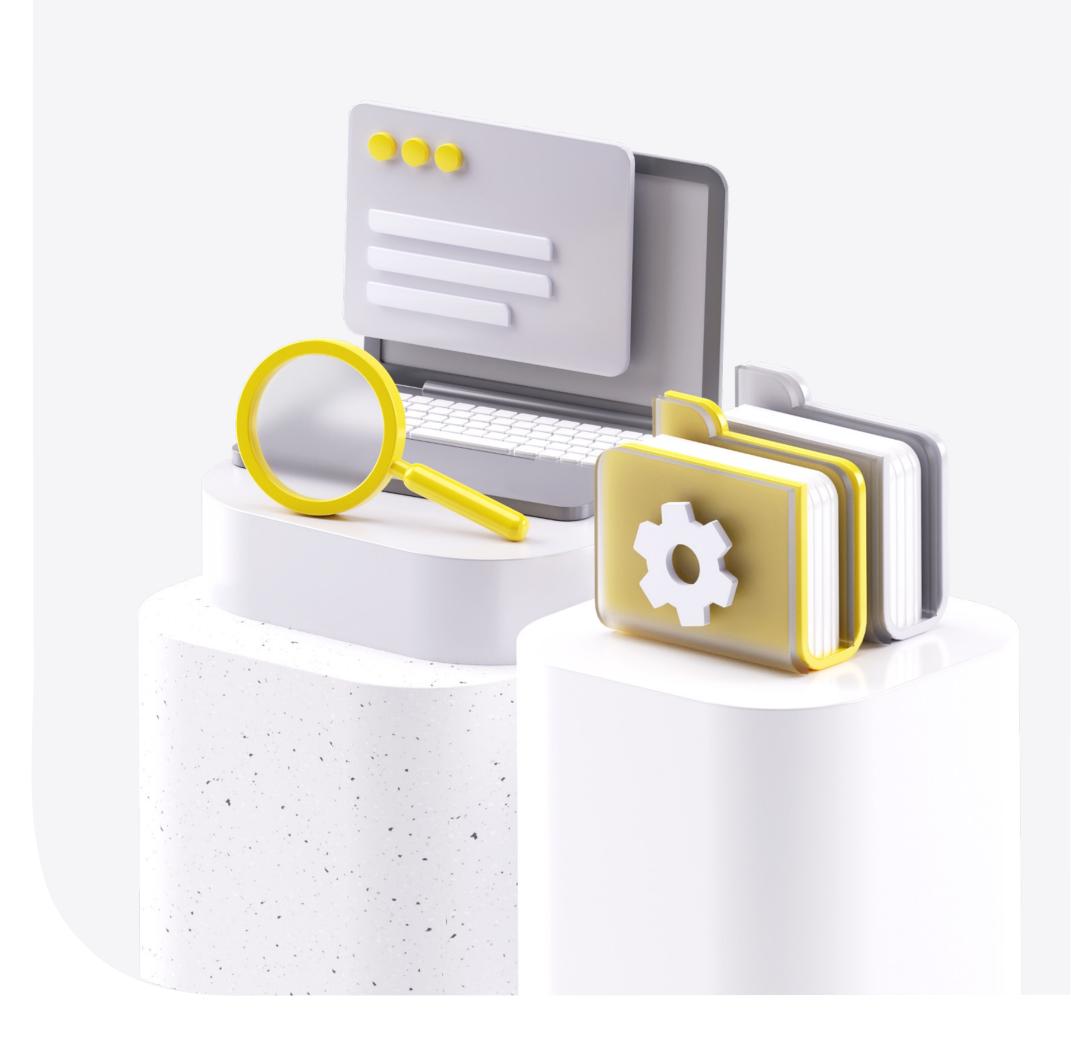




Data Detective

Платформа для поиска и каталогизации данных





Александр Долгих

Product Owner Data Detective



@sasha_dolgikh



a.i.dolgikh@tinkoff.ru



Роман Митасов

Глава Back-end разработки



@DrankUser

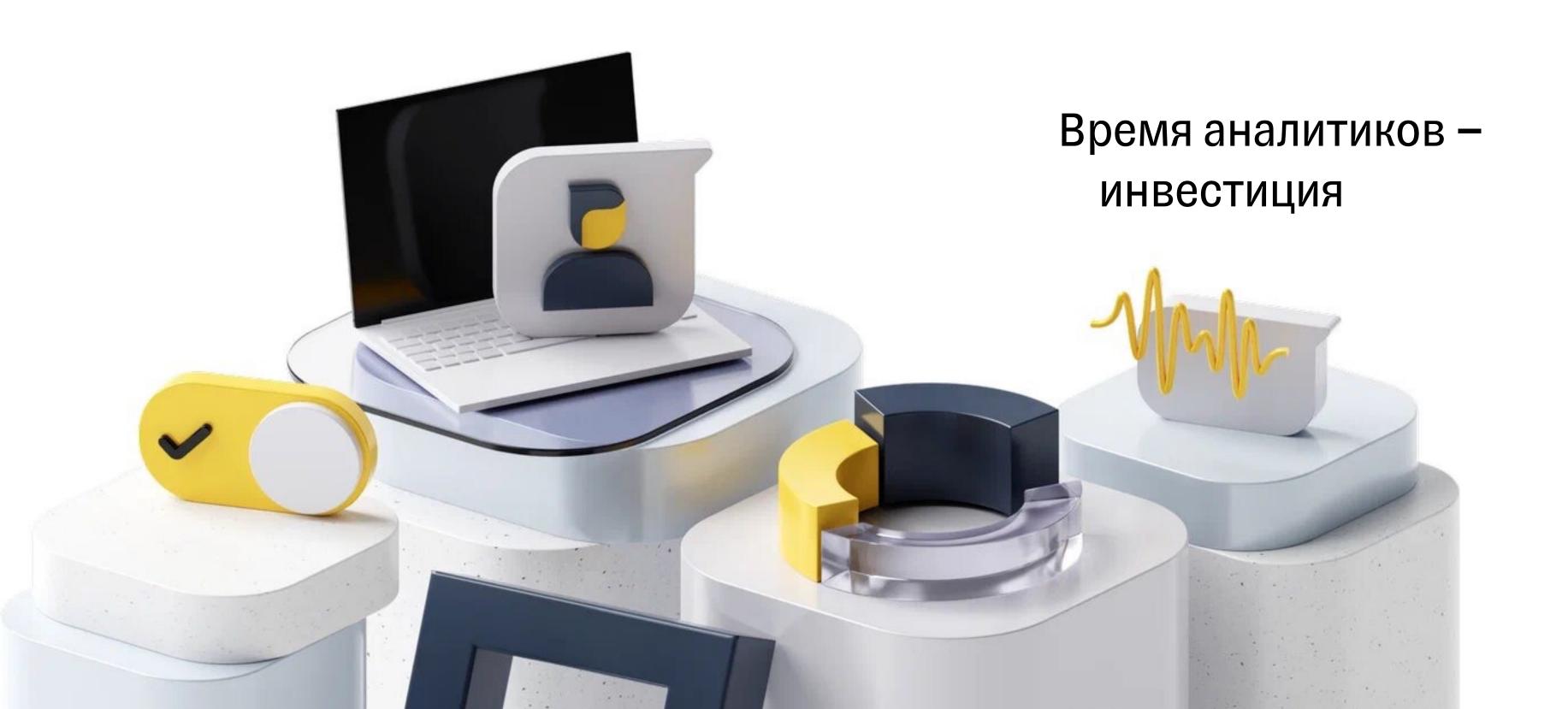


r.mitasov@tinkoff.ru



Зачем нужна каталогизация?

Данные – ценнейший актив компании



Метаданные – знания о данных





Масштаб проблемы у нас

6000 +

Пользователей данных 25+

Кластеров СУБД

150 000 +

Таблиц

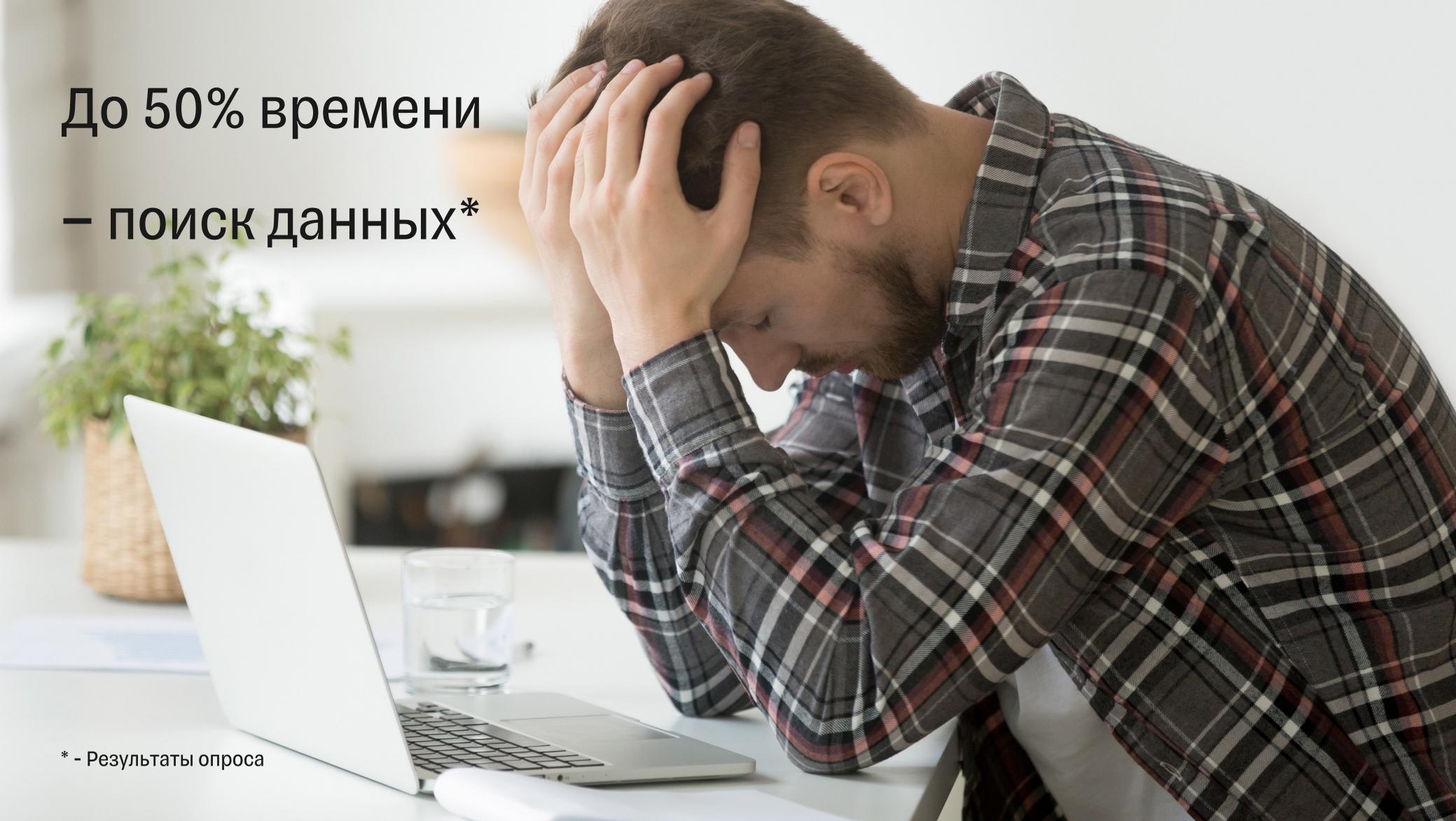
9000 +

Отчетов BI

38 000 +

Self-Service Zeppelin notes 14 000 +

ETL pipelines





Критерии



Гибкая модель метаданных

Большое количество внутренних ИТ-решений



Детальный Data Lineage

Разобраться в наших потоках данных – очень нетривиальная задача



Удобный UI

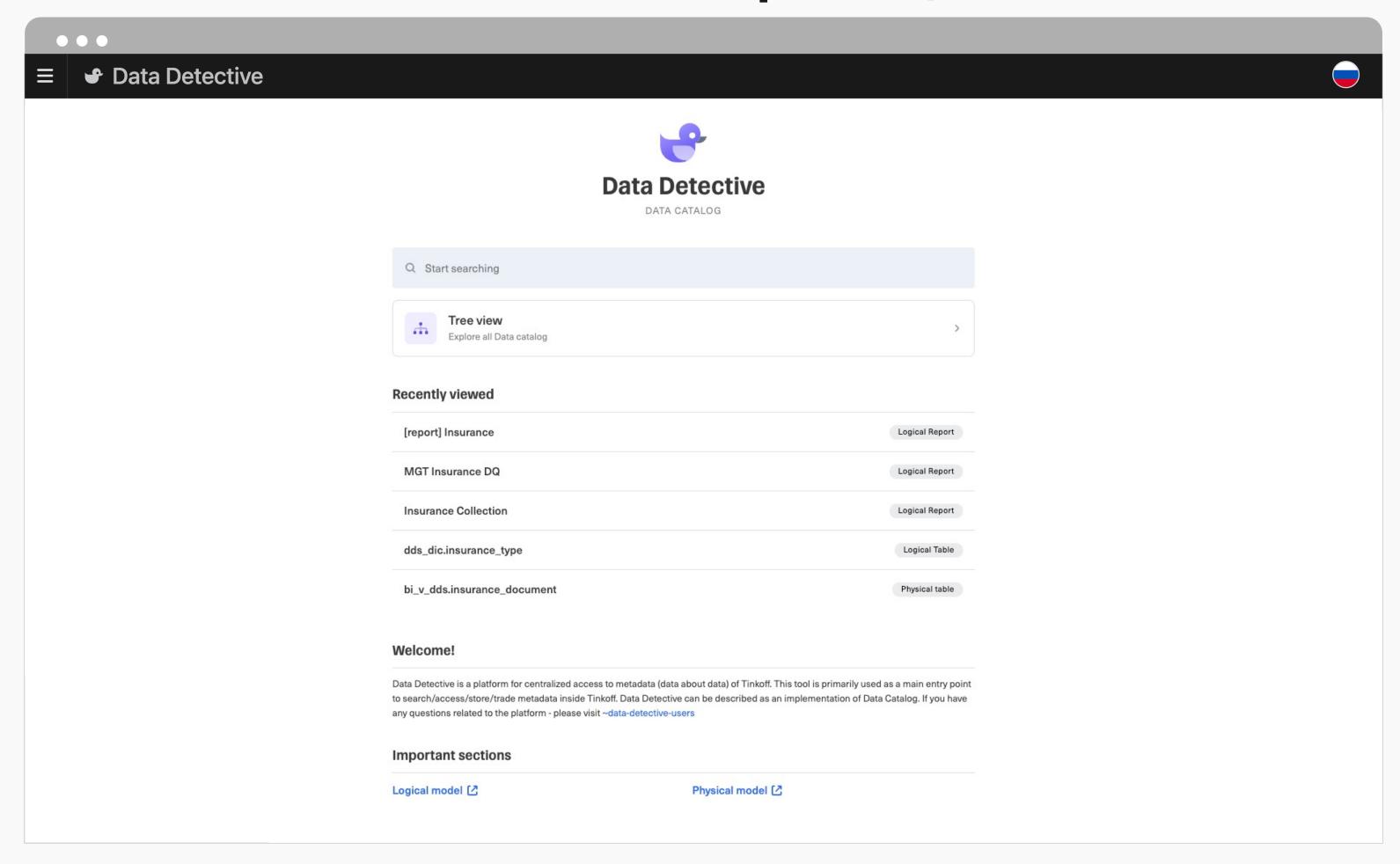
Пользователи не должны страдать



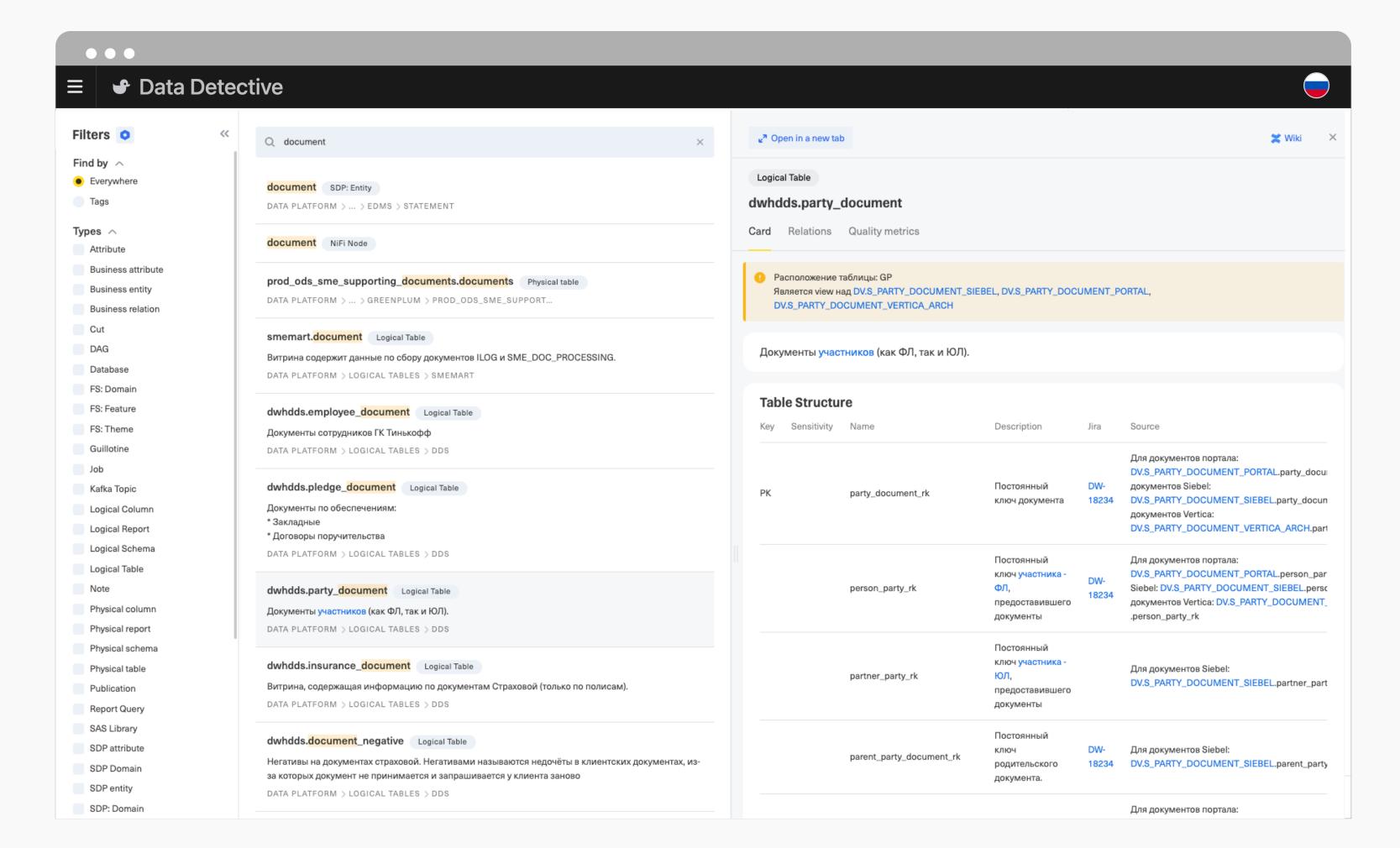
Стоимость

Не хотим тратить слишком много денег

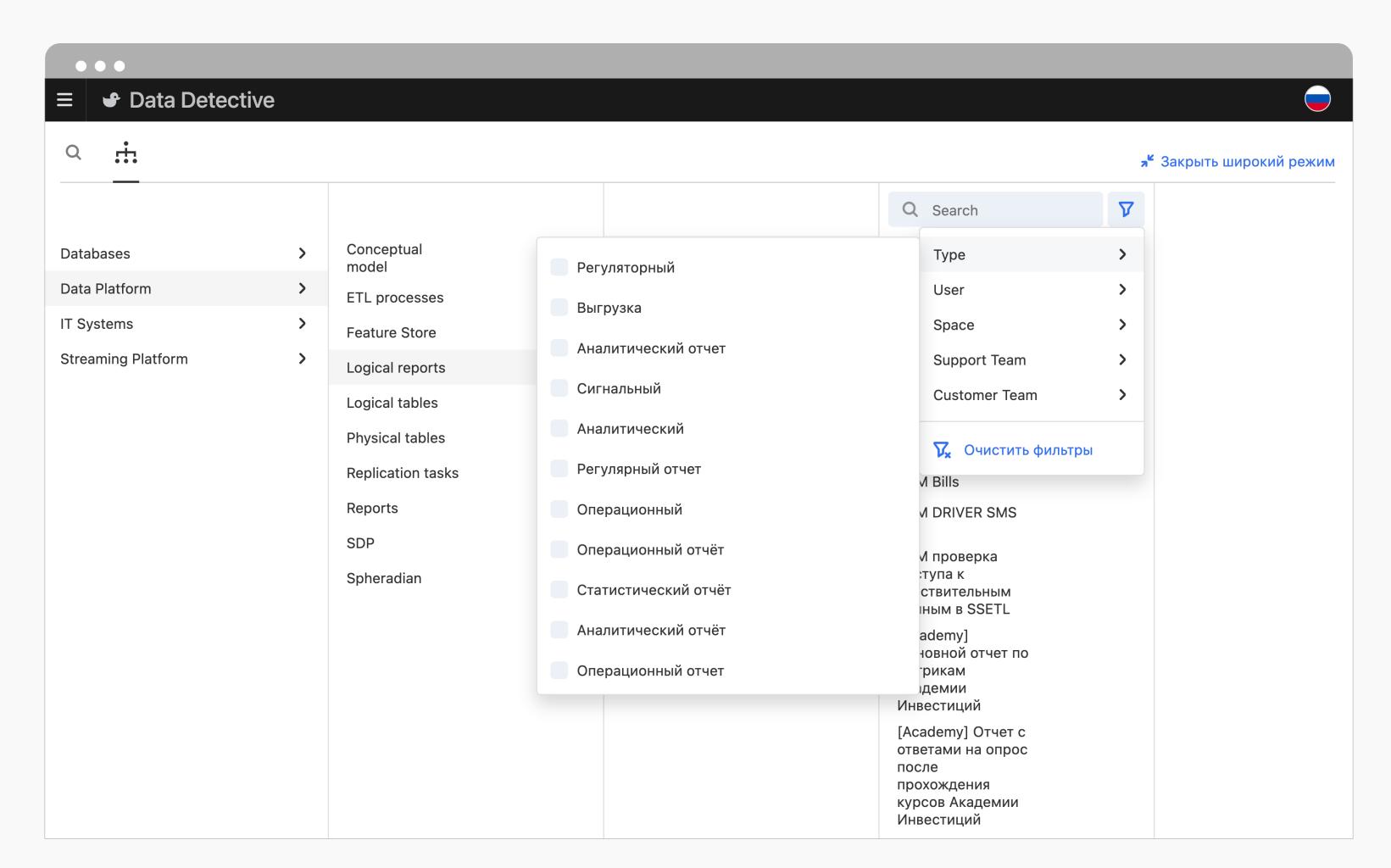
Главная страница



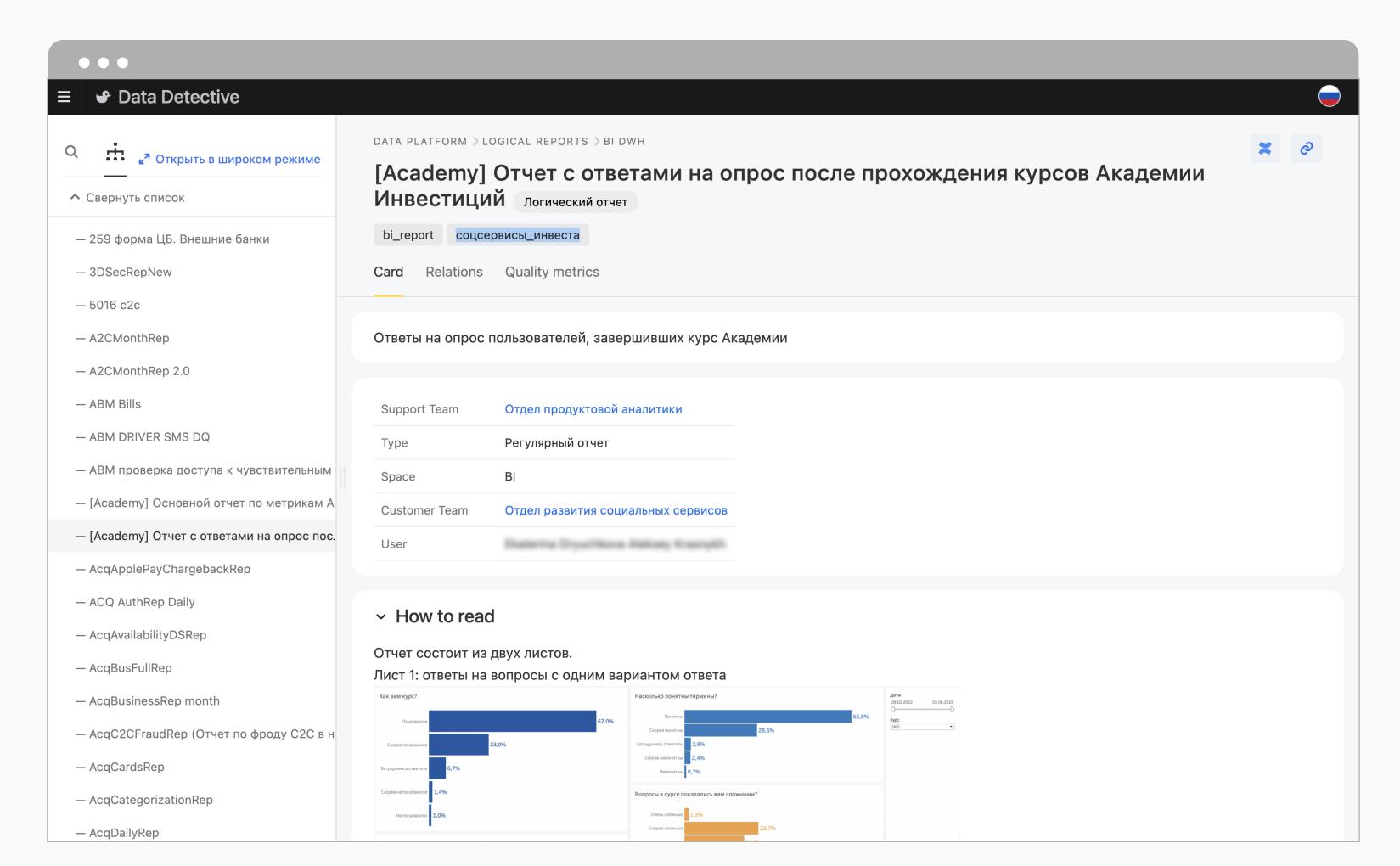
Полноэкранный поиск



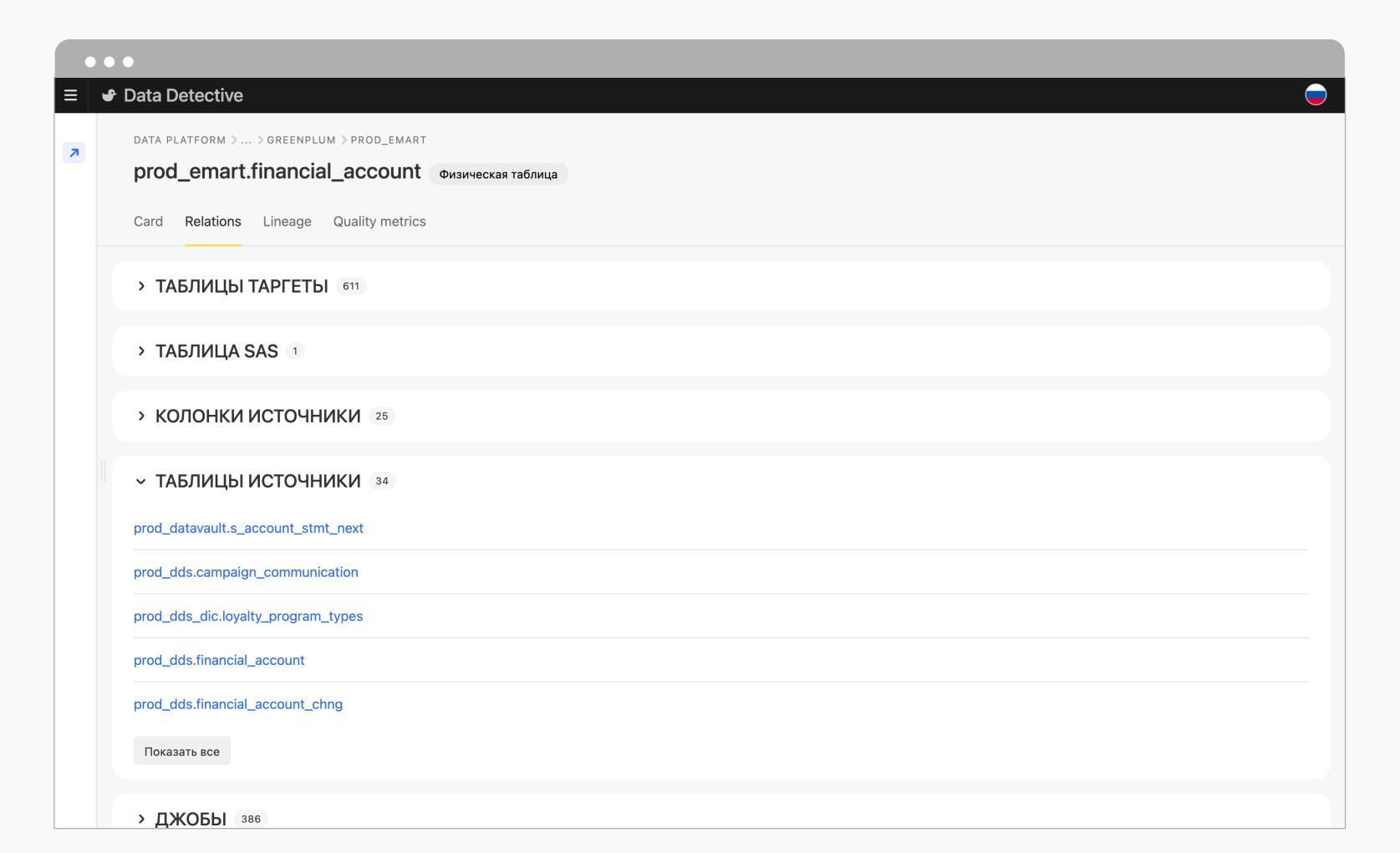
Полноэкранное дерево объектов



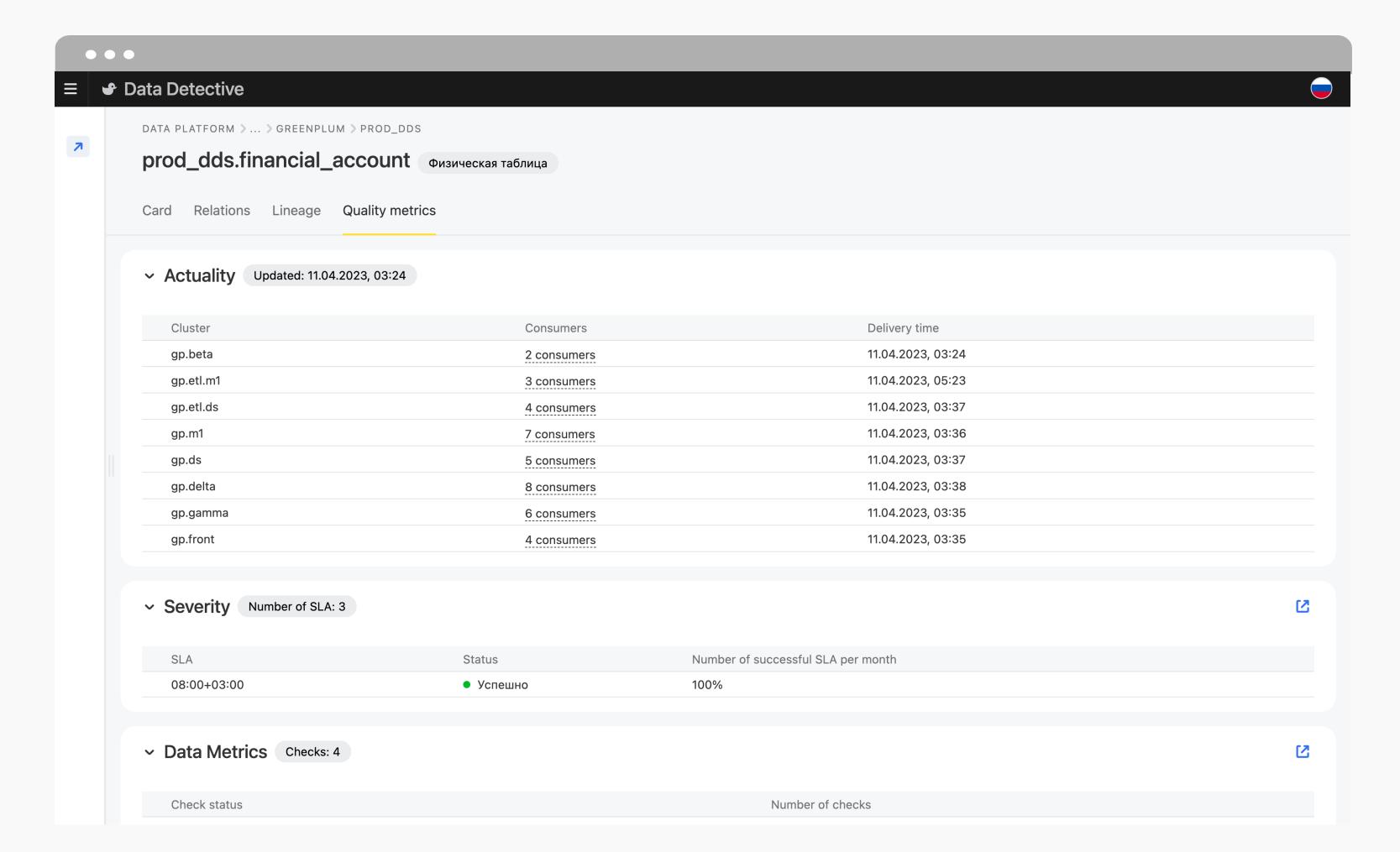
Карточка



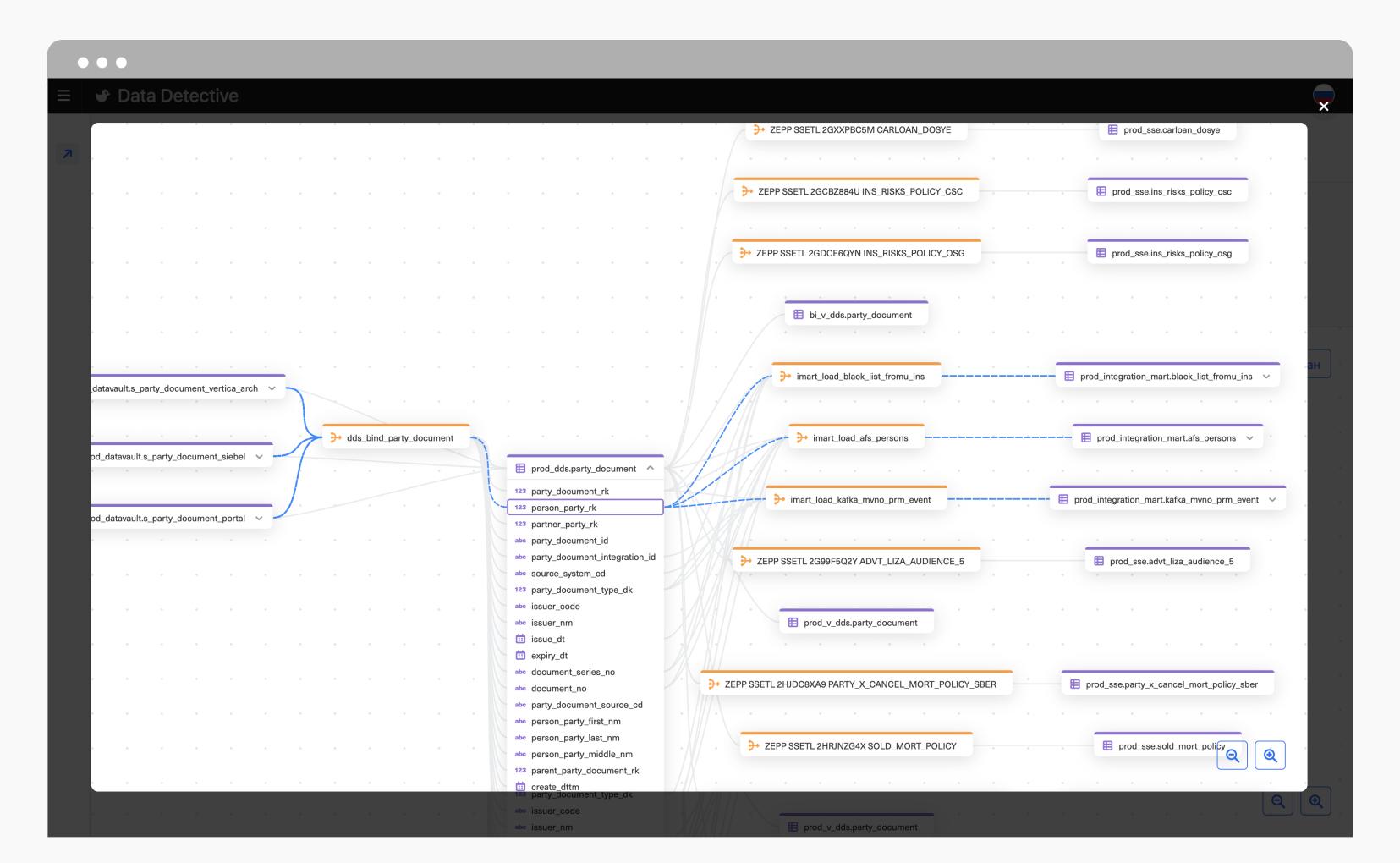
Связи



Метрики качества данных



Data Lineage



Технические особенности



Архитектура Data Detective

Tinkoff















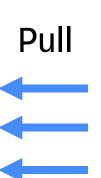




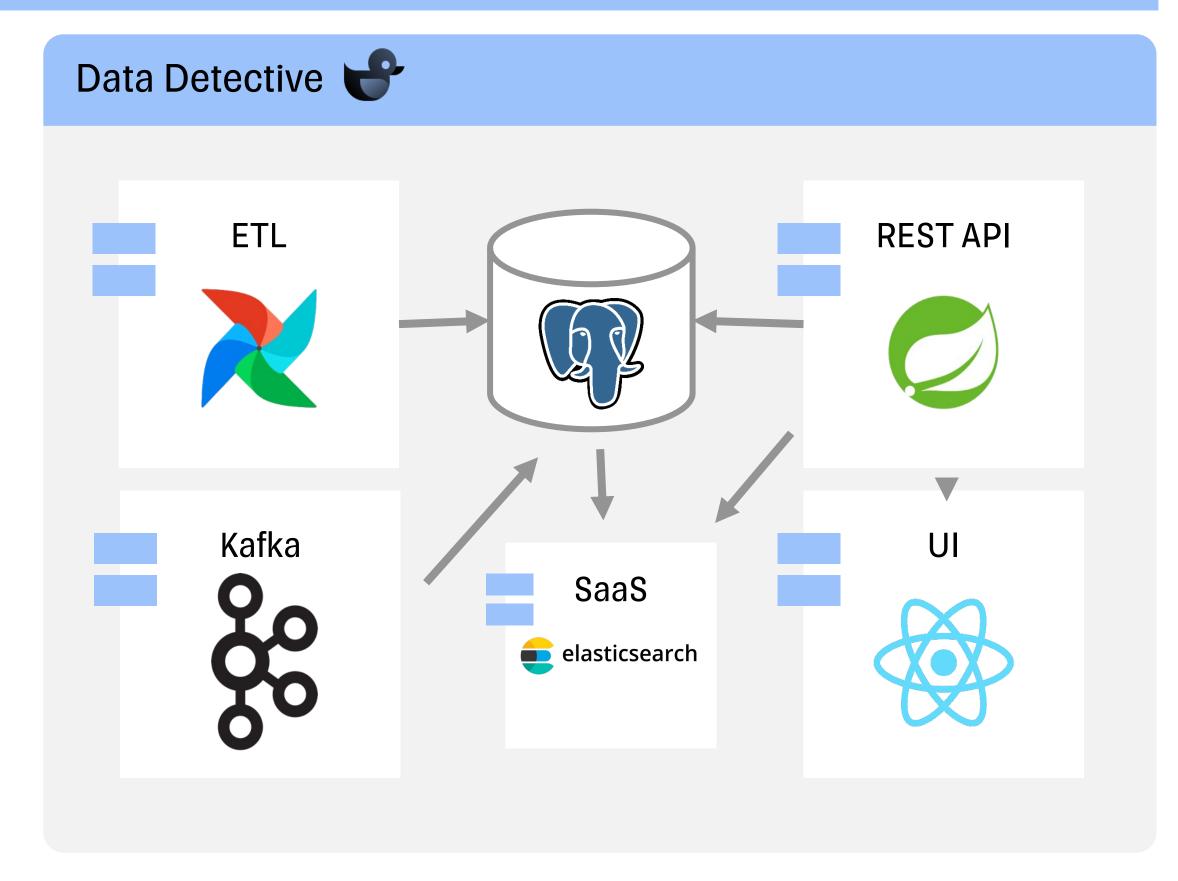












Модель данных Data Detective

Таблицы

Entity – сущности

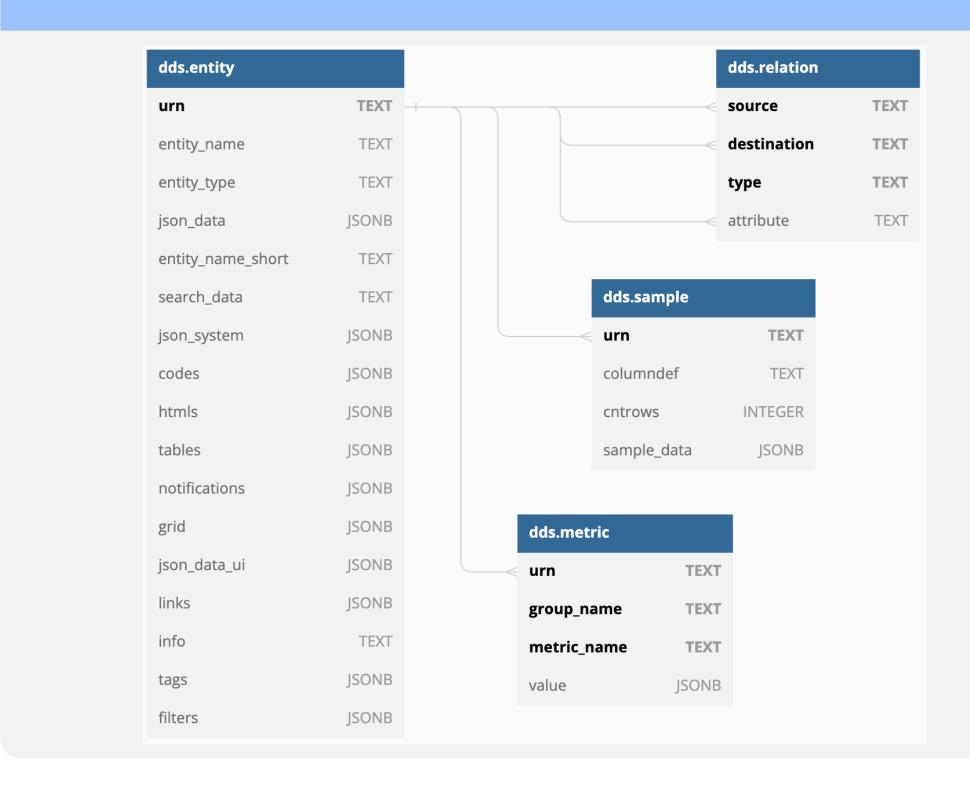
Relation – связи между сущностями

Sample – примеры данных

Metric – загруженные метрики

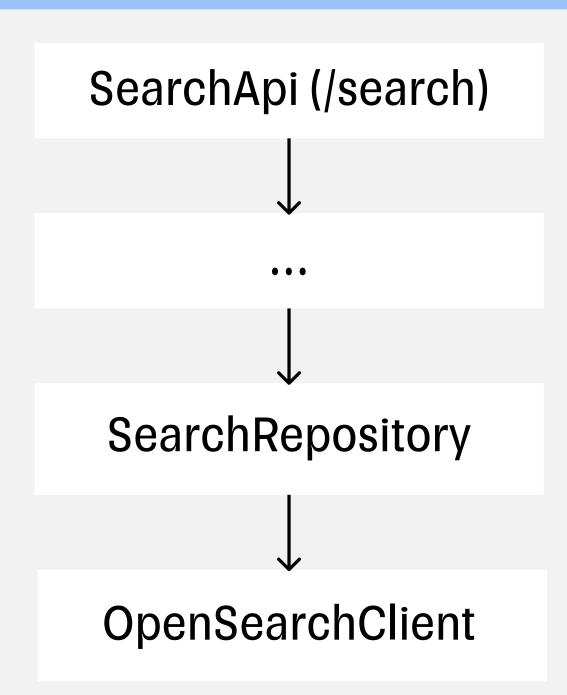
URN – уникальное имя сущности

Data Detective DB Model



Расширяемая архитектура кода

DD API



DD API

SearchApi (/search)
....

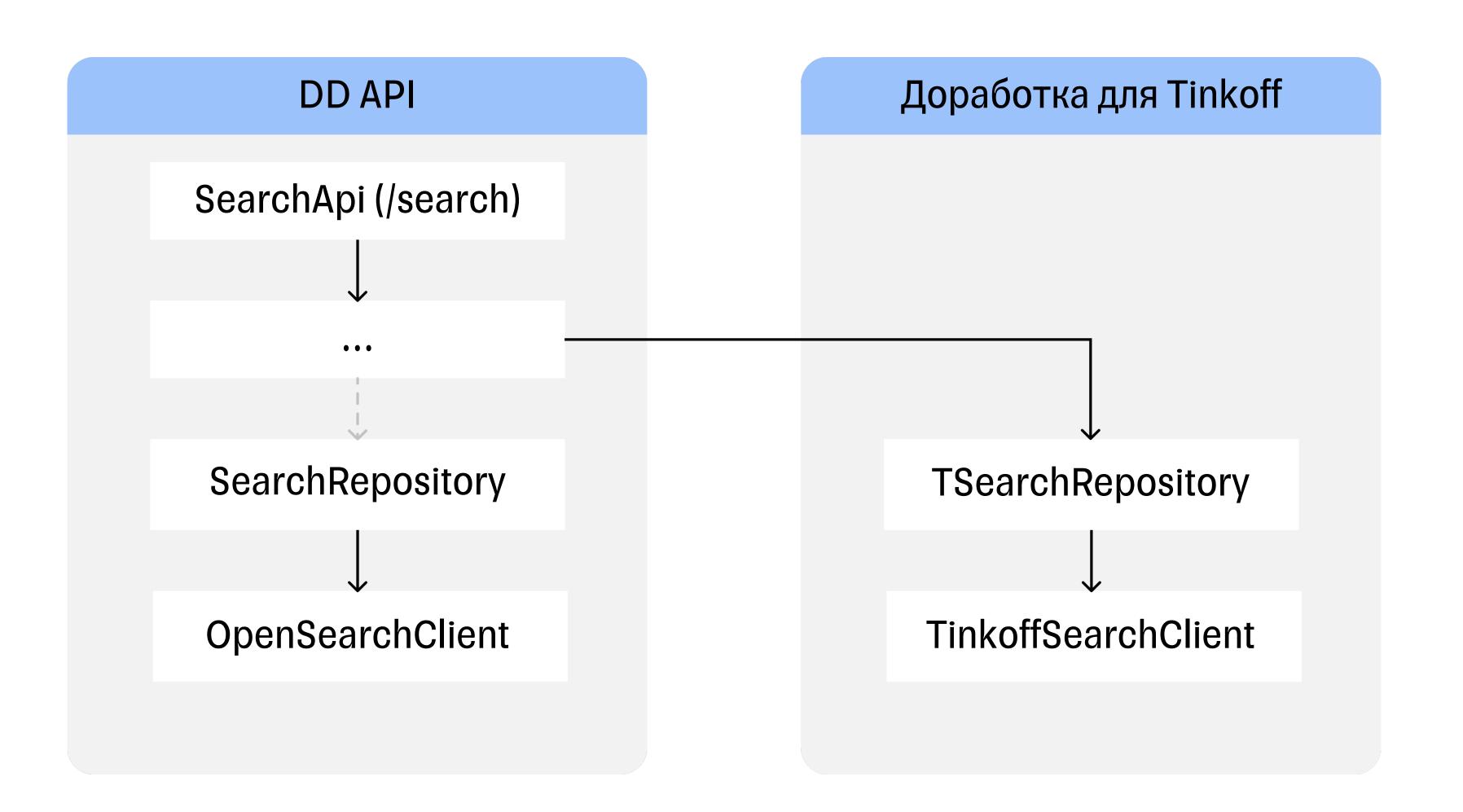
OpenSearchClient

SearchRepository

Доработка для Tinkoff

TSearchRepository

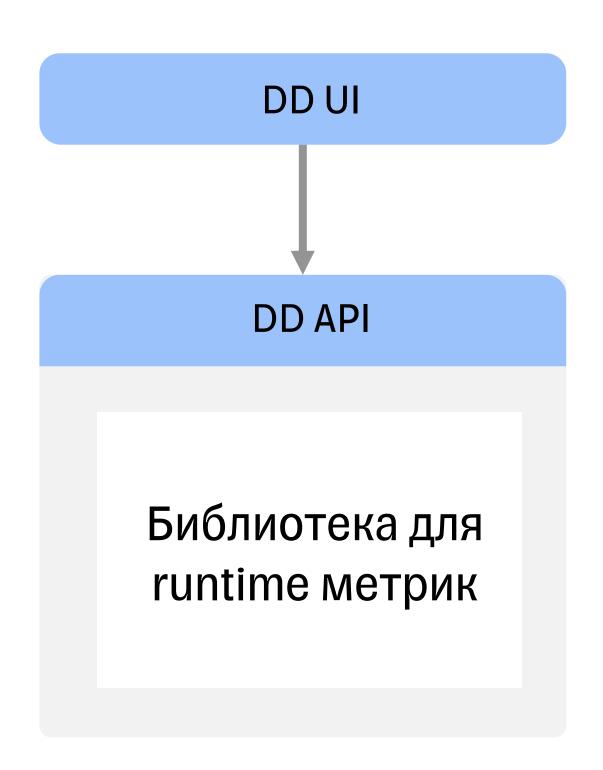
TinkoffSearchClient



Real-time метрики



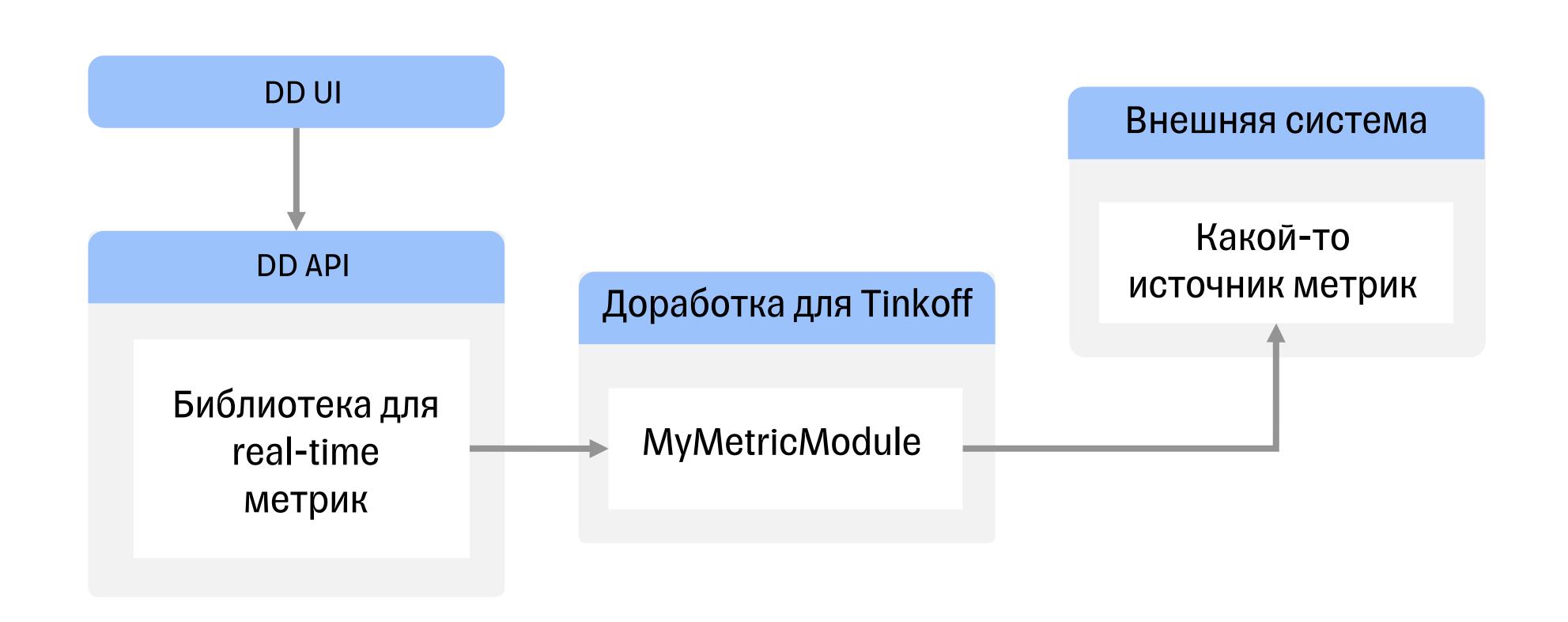
Real-time метрики



Внешняя система

Какой-то источник метрик

Real-time метрики





Кейсы в Тинькофф

Основной кейс

Упростили поиск данных в компании

Задача

Разработать быстрый, удобный и устойчивый к нагрузкам инструмент для поиска данных

Решение

Сделали шаблон, по которому описываются объекты данных. Затем привели все описания к новому шаблону

Загрузили описания в Data Detective. Научились считать индекс качества описаний и начали его поднимать

Научили пользователей описывать их объекты данных Результат

1200

пользователей еженедельно обращаются к поиску

NPS

7,9

CTR

72%

Реализовали миграцию более 5000 процессов

Задача

Спланировать и бесшовно реализовать миграцию более 5000 процессов обработки данных со старого инструмента на новый

Решение

С помощью Data Detective определили четкий объем работ и контролировали ход проекта благодаря наличию метаданных по всем процессам

Результат

> 5000

процессов бесшовно перевели на новый инструмент

0%

пользователей заметили миграцию

Создали карту чувствительных данных компании

Задача

Найти и разметить чувствительные данные в огромной экосистеме данных компании

Решение

С помощью Data Detective предоставили метаданные, которые распарсили для поиска нужной информации и разметки чувствительных данных

Научили платформу автоматически следить за тем, какие данные поступают в компанию, и определять их чувствительность

Результат

Создали автоматический процесс поиска и разметки чувствительных данных. Нашли и разметили:

1000

таблиц

4700

полей

Если хочется подробнее



Канал продукта

t.me/datadetective



data-detective.dev