Контроль качества данных с помощью

Arenadata Catalog Data Quality Framework (ADC.DQF)

Юрий Горынцев
Руководитель отдела модуля DQ

Arenadata Catalog

ADC Data Quality Framework

ADC.DQF - высокопроизводительный сервис выполнения проверок **качества** данных
по настраиваемым **правилам** с консолидацией **результатов** и формированием **инцидентов** качества

- Соблюдение критериев качества данных
- Исполнение требований регуляторов
- Поддержка процессов импортозамещения и цифровизации с применением методик Data Quality на новых платформах
- Миграция, консолидация данных, предобработка данных
- Проактивное реагирование на инциденты качества и аномалии в данных
- Централизация проверок, реализация методологий Data Quality и Data Governance в рамках одного интерфейса

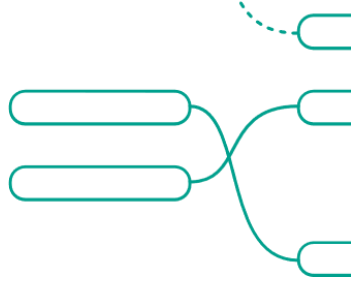
Arenadata Catalog

Преимущества ADC.DQF (SQL проверки)

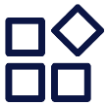
Централизация проверок качества данных в ADC

SQL (Push-Down)	DQF (Push-Up)
Уровень СУБД	Уровень клиента, бэкенда и СУБД
Нагрузка носит пиковый характер и может повредить проду	Возможность **управления нагрузкой**
Является **последним рубежом** обороны. Защищает от ошибок в коде приложения и прямого доступа к БД.	Предоставляет функционал для **дополнительных «барьеров» (Firewall)** на пути невалидных данных.
Сложно реализовать очень сложную бизнес-логику. Код на SQL/процедурах для сложной логики может быть тяжелым для поддержки. Нет интеграции по API	Позволяет **реализовать с минимальными трудозатратами** любую сложную логику используя собственный язык DSL (включая кросс системные сверки) Есть интеграция
Сложнее при сопровождении: Миграции БД, версионирование, тесты обычно сложнее писать и запускать.	**Проще при сопровождении:** Собственный графический редактор с функционалом отладки проверок качества данных.
Сильная связность с СУБД: Смена БД(например, с PostgreSQL на MongoDB) может потребовать полного переписывания скрипта и изменения логики валидации.	**Слабая связность**: Логика проверок не зависит от конкретной СУБД.

Несколько точек контроля



Клиентское приложение



Backend

OLTP СУБД

```
{
  "firstName": "Gtnh",
  "lastName": "Иванов",
  "birthDate": {
    "year": "2008",
    "month": "05",
    "day": null
  }
}
```

id	firstName	lastName	bd_year	bd_month	bd_day
1	Сергей		1999	02	12
2	Петр	Петров	1990	11	11
3	Василий	Васильев		06	09

Arenadata Catalog

Возможности проверок

Виды проверок

- Форматно-логический контроль
- Проверки внутри модели
- Интеграционные проверки

YAML-конфигурация
Гибкая настройка правил через синтаксис YAML 1.2.2

Алгоритмы проверок:

- Строковые значения
- Даты
- Арифметические операции
- Условные операторы
- Алгоритмические проверки

Расширенные алгоритмы:

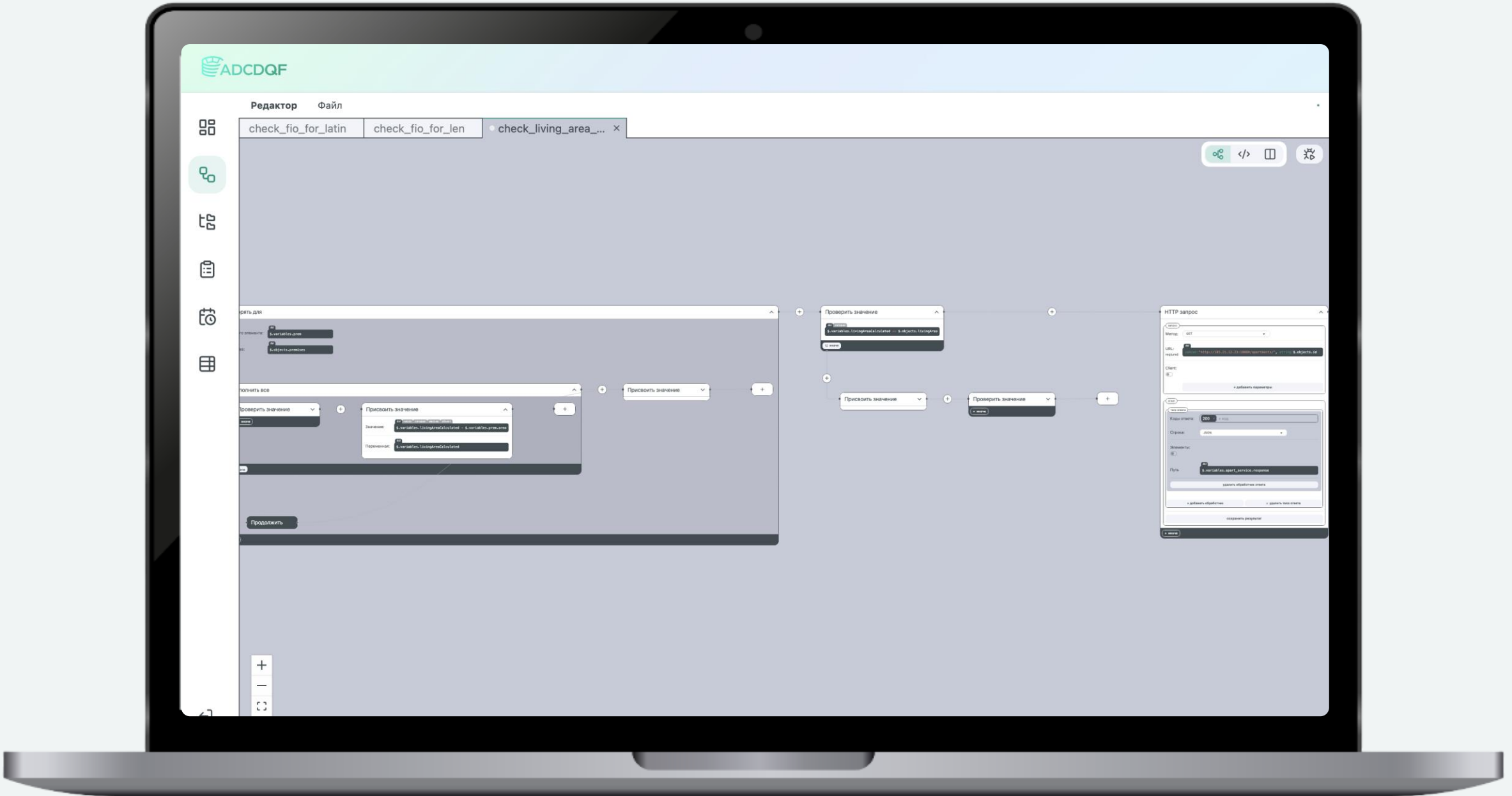
- СНИЛС, ИНН, ОГРН
- HTTP-, SOAP-, GraphQL-, JDBC- запросы

Циклы:

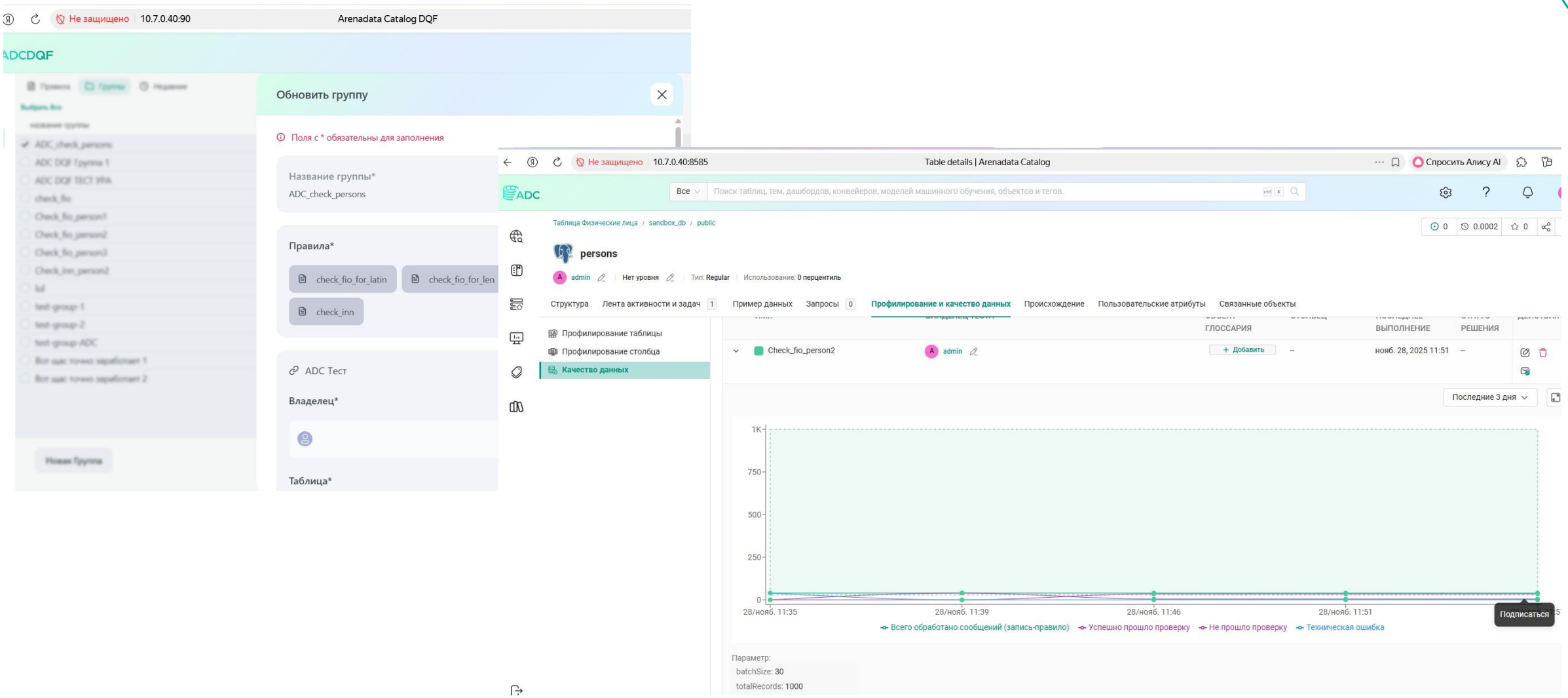
- Для каждого элемента коллекции
- Пока выполняется условие

Arenadata Catalog

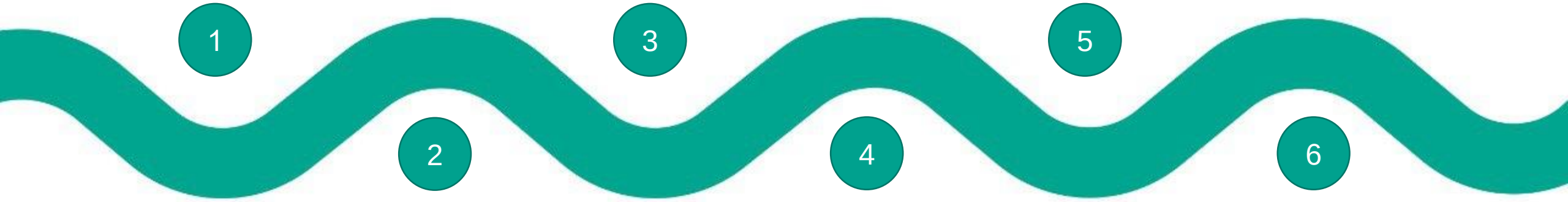
Интерфейс редактора



Интеграция тестов DQF в ADC



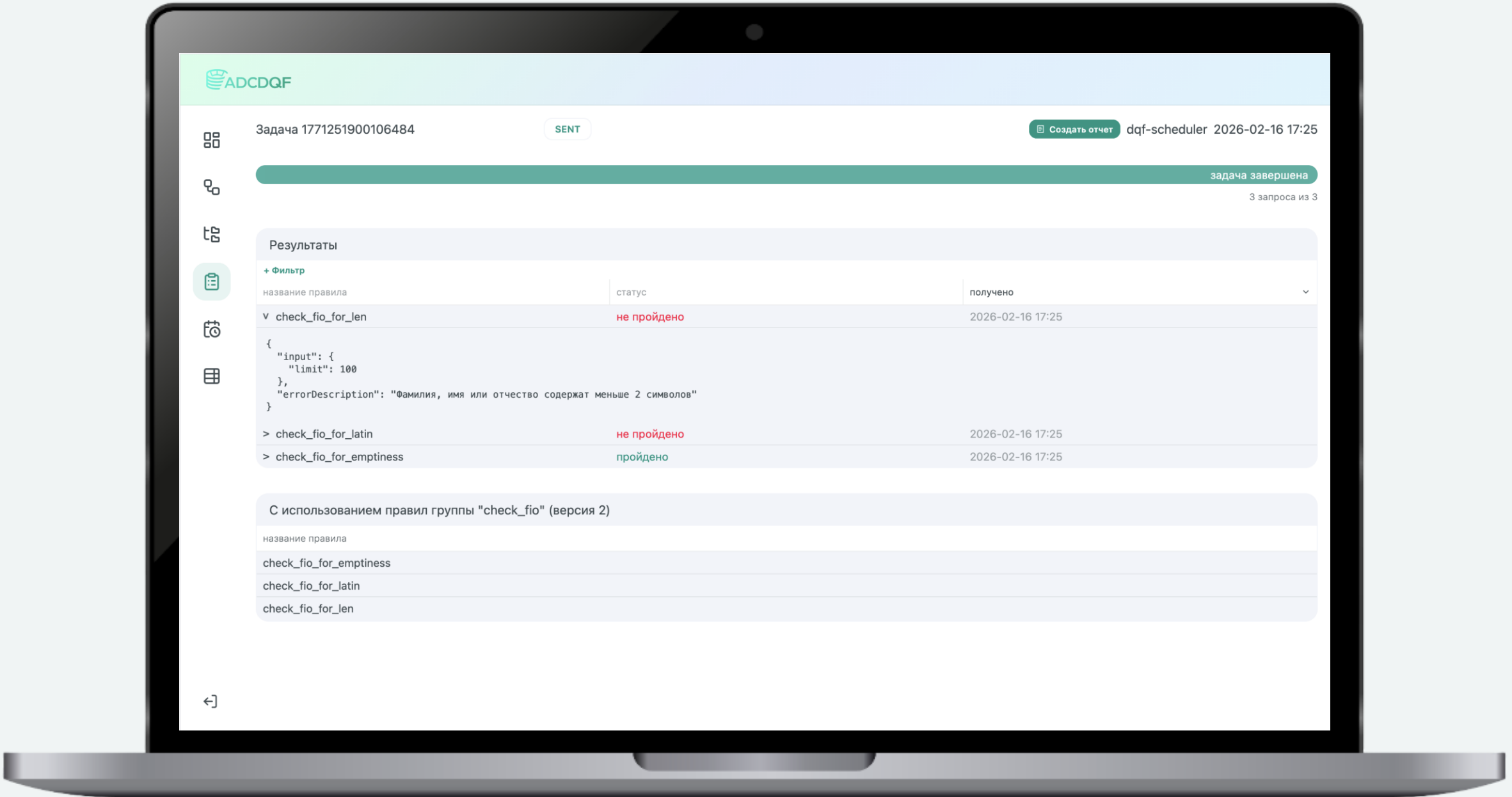
Интегрированный процесс



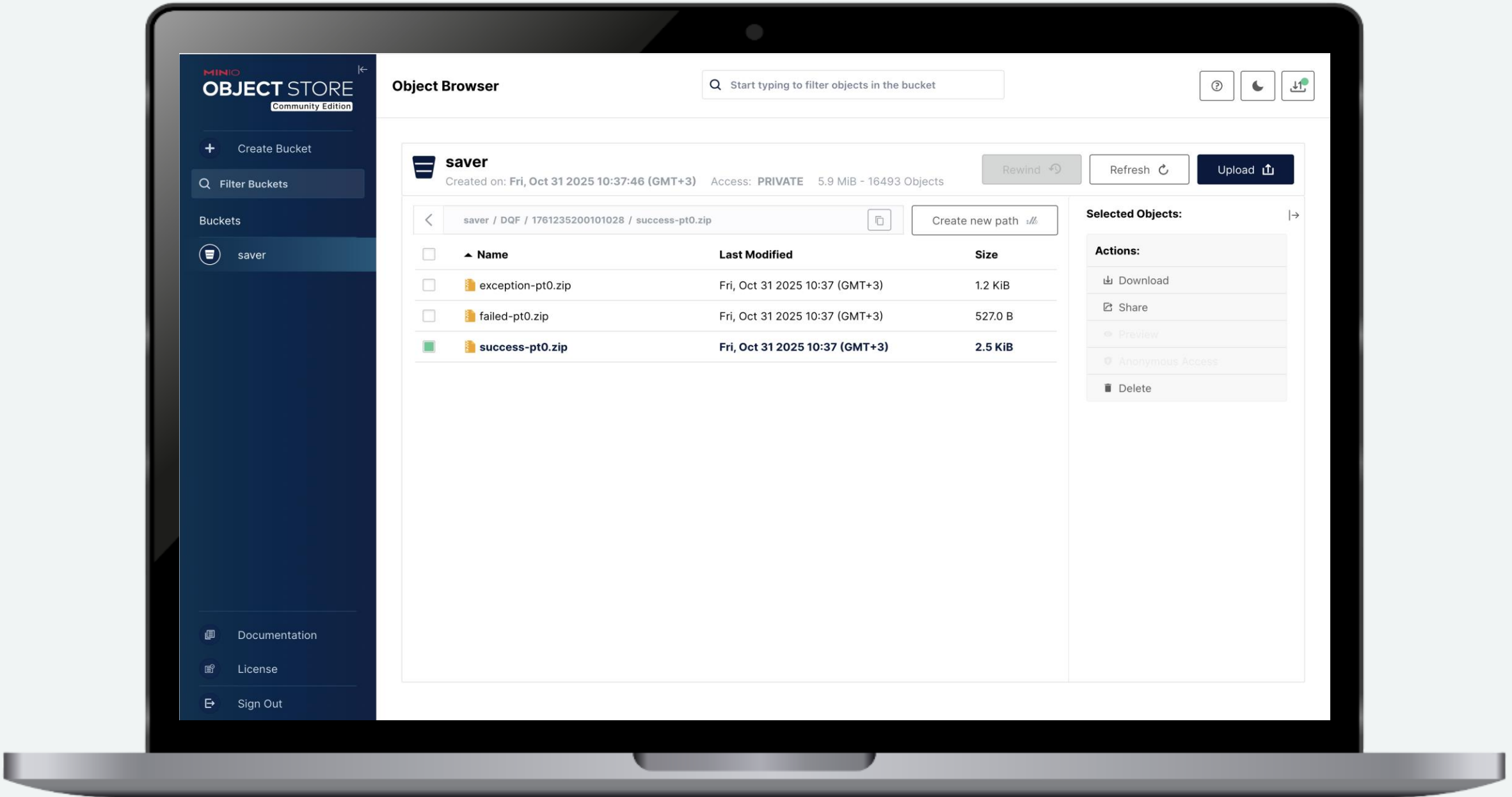
1. Создание бизнес-требования в ADC
2. Создание правила в ADC.DQF
3. Связь правила и бизнес-требования в ADC
4. Согласование правила в ADC
5. Настройка и исполнение проверки
6. Анализ результатов

Arenadata Catalog

Отображение результатов задач



Сохранение всех результатов



Дальнейшие планы

1.0.1

Сохранение результатов в CSV/XLS

- Короткий семпл в человекочитаемом виде
- Нотификация об ухудшении качества сразу с примером данных

Теги правил и групп

- Улучшение навигации в каталоге правил
- Дополнительные измерения для группировки правил

1.0.2

Улучшение языка выражений

- Замена по шаблону
- Явное указание выходных данных проверки

Шаблонизация правил

- Отвязка от физического представления
- Решение проблемы большого количества сходных правил

Метаданные в правилах

- Объединение базовых алгоритмов в смысловые блоки
- Комментарии

Дальнейшие планы

1.0.1

Сохранение результатов в CSV/XLS

- Короткий семпл в человекочитаемом виде
- Нотификация об ухудшении качества сразу с примером данных

Теги правил и групп

- Улучшение навигации в каталоге правил
- Дополнительные измерения для группировки правил

1.0.2

Улучшение языка выражений

- Замена по шаблону
- Явное указание выходных данных проверки

Шаблонизация правил

- **Отвязка от физического представления**
- **Решение проблемы большого количества сходных правил**

Метаданные в правилах

- Объединение базовых алгоритмов в смысловые блоки
- Комментарии

Дальнейшие планы

1.0.3

Сервис подготовки данных для DQ

- Кеширование для использования в нескольких проверках
- Трансформации для выполнения проверок

Использование результатов профилирования в проверках

- Превентивный поиск аномалий
- Сравнение данных с ожиданиями

1.1.0

Формирование инцидентов качества

- Настраиваемый воркфлоу
- Отображение критериев на графиках

Использование AI-DS для правил

- Помощь при разработке правил
- Массовое создание правил

Отладка правил

- Пошаговый запуск правил
- Анализ промежуточных результатов

Arenadata Catalog

Дальнейшие планы

1.0.3

Сервис подготовки данных для DQ

- Кеширование для использования в нескольких проверках
- Трансформации для выполнения проверок

Использование результатов профилирования в проверках

- Превентивный поиск аномалий
- Сравнение данных с ожиданиями

1.1.0

Формирование инцидентов качества

- Настраиваемый воркфлоу
- Отображение критериев на графиках

Использование AI-DS для правил

- **Помощь при разработке правил**
- **Массовое создание правил**

Отладка правил

- Пошаговый запуск правил
- Анализ промежуточных результатов

Arenadata Catalog