

Контроль качества данных с помощью

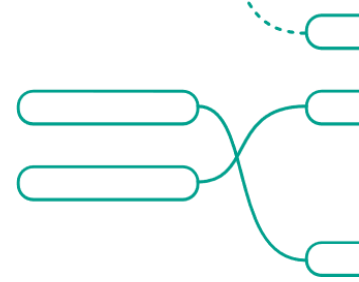
Arenadata Catalog Data Quality Framework (ADC.DQF)

Юрий Горынцев

Руководитель отдела модуля DQ



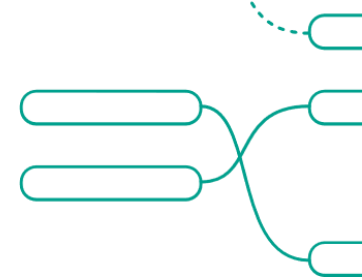
ADC Data Quality Framework



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

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 продукту	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

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



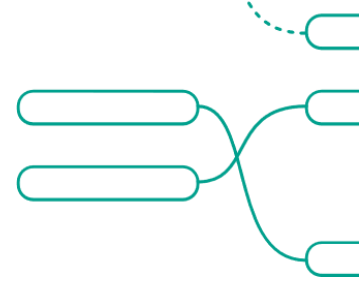
Backend



OLTP СУБД

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

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

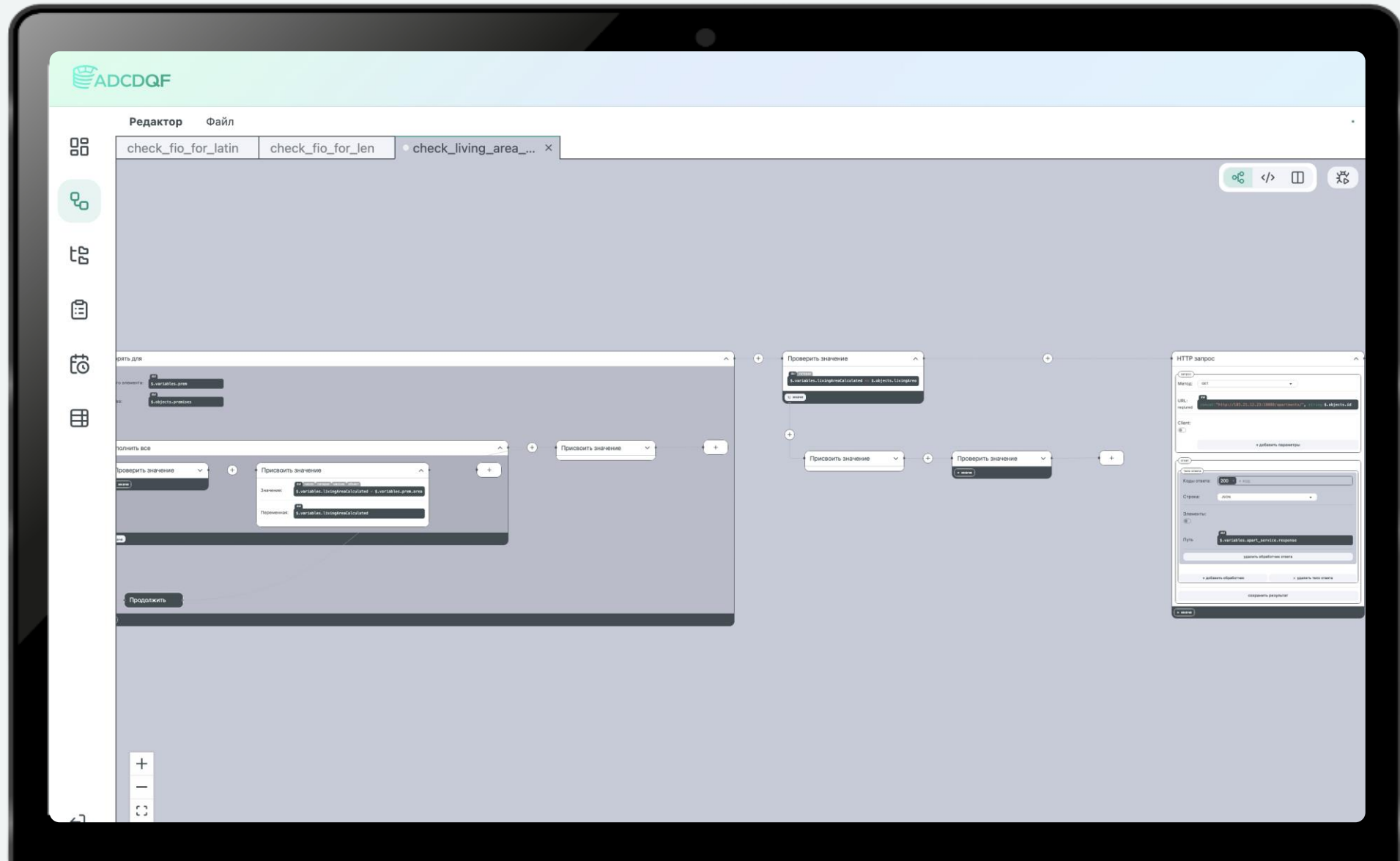
Расширенные алгоритмы:

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

Циклы:

- Для каждого элемента коллекции
- Пока выполняется условие

Интерфейс редактора



Интеграция тестов DQF в ADC



Arenadata Catalog DQF

Обновить группу

Поля с * обязательны для заполнения

Название группы*
ADC_check_persons

Правила*

- check_fio_for_latin
- check_fio_for_len
- check_inn

ADC Тест

Владелец*

Таблица*

Table details | Arenadata Catalog

ADC

Все

Поиск таблиц, тем, дашбордов, конвейеров, моделей машинного обучения, объектов и тегов.

Таблица Физические лица / sandbox_db / public

persons

admin | Нет уровня | Тип: Regular | Использование: 0 перцентиль

Структура | Лента активности и задач | Пример данных | Запросы | 0 | Профилирование и качество данных | Происхождение | Пользовательские атрибуты | Связанные объекты

Профилирование таблицы

Профилирование столбца

Качество данных

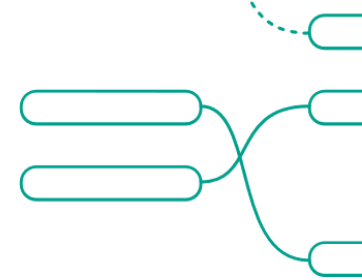
ГЛОССАРИЯ	ВЫПОЛНЕНИЕ	РЕШЕНИЯ
+ Добавить	нояб. 28, 2025 11:51	

Последние 3 дня

Подписаться

Параметр:
batchSize: 30
totalRecords: 1000

Интегрированный процесс



Создание бизнес-
требования в ADC

1

Связь правила
и бизнес-требования в ADC

3

Настройка
и исполнение проверки

5

Создание правила
в ADC.DQF

2

Согласование
правила в ADC

4

Анализ результатов

6

Отображение результатов задач

The screenshot displays the ADCDQF interface for a task. At the top, the task ID is 'Задача 1771251900106484' with a 'SENT' status. A 'Создать отчет' button and the scheduler name 'dqf-scheduler' with the time '2026-02-16 17:25' are also visible. A green progress bar indicates 'задача завершена' (task completed) with '3 запроса из 3' (3 requests of 3).

The 'Результаты' (Results) section contains a table with the following data:

название правила	статус	получено
v check_fio_for_len	не пройдено	2026-02-16 17:25
> check_fio_for_latin	не пройдено	2026-02-16 17:25
> check_fio_for_emptiness	пройдено	2026-02-16 17:25

The failed rule 'v check_fio_for_len' has a detailed error description:

```
{ "input": { "limit": 100 }, "errorDescription": "Фамилия, имя или отчество содержат меньше 2 символов" }
```

Below the table, a section titled 'С использованием правил группы "check_fio" (версия 2)' lists the rules used: 'check_fio_for_emptiness', 'check_fio_for_latin', and 'check_fio_for_len'.

Сохранение всех результатов

The screenshot displays the MINIO Object Store interface. On the left is a dark blue sidebar with the 'OBJECT STORE Community Edition' logo and navigation options: '+ Create Bucket', 'Filter Buckets', a 'Buckets' list containing 'saver', and links for 'Documentation', 'License', and 'Sign Out'. The main area is titled 'Object Browser' and features a search bar with the placeholder 'Start typing to filter objects in the bucket'. Below the search bar, the bucket 'saver' is shown with details: 'Created on: Fri, Oct 31 2025 10:37:46 (GMT+3)', 'Access: PRIVATE', and '5.9 MiB - 16493 Objects'. Action buttons for 'Rewind', 'Refresh', and 'Upload' are visible. The current path is 'saver / DQF / 1761235200101028 / success-pt0.zip'. A table lists the objects in the bucket:

<input type="checkbox"/>	Name	Last Modified	Size
<input type="checkbox"/>	exception-pt0.zip	Fri, Oct 31 2025 10:37 (GMT+3)	1.2 KiB
<input type="checkbox"/>	failed-pt0.zip	Fri, Oct 31 2025 10:37 (GMT+3)	527.0 B
<input checked="" type="checkbox"/>	success-pt0.zip	Fri, Oct 31 2025 10:37 (GMT+3)	2.5 KiB

On the right side, the 'Selected Objects:' panel shows a list of actions for the selected object: 'Download', 'Share', 'Preview', 'Anonymous Access', and 'Delete'.

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

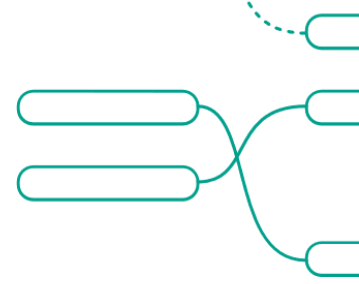
- Замена по шаблону
- Явное указание выходных данных проверки

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

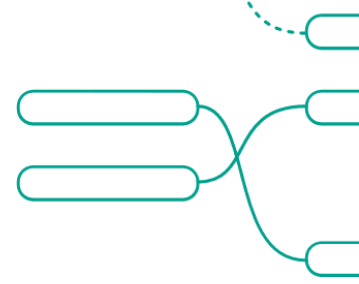
- Замена по шаблону
- Явное указание выходных данных проверки

Шаблонизация правил

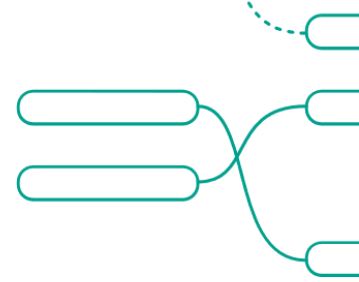
- **Отвязка от физического представления**
- **Решение проблемы большого количества сходных правил**

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

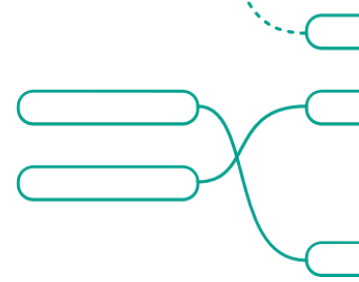
Использование AI-DS для правил

- Помощь при разработке правил
- Массовое создание правил

Отладка правил

- Пошаговый запуск правил
- Анализ промежуточных результатов

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