



Качество данных в современном цифровом банке

Руслан Хохлов
Департамент управления данными

gazprombank.ru





Система развития искусственного интеллекта.

Комплекс подходов, практик и процессов разработки и применения технологий ИИ для решения конкретных задач организации

Система управления данными.

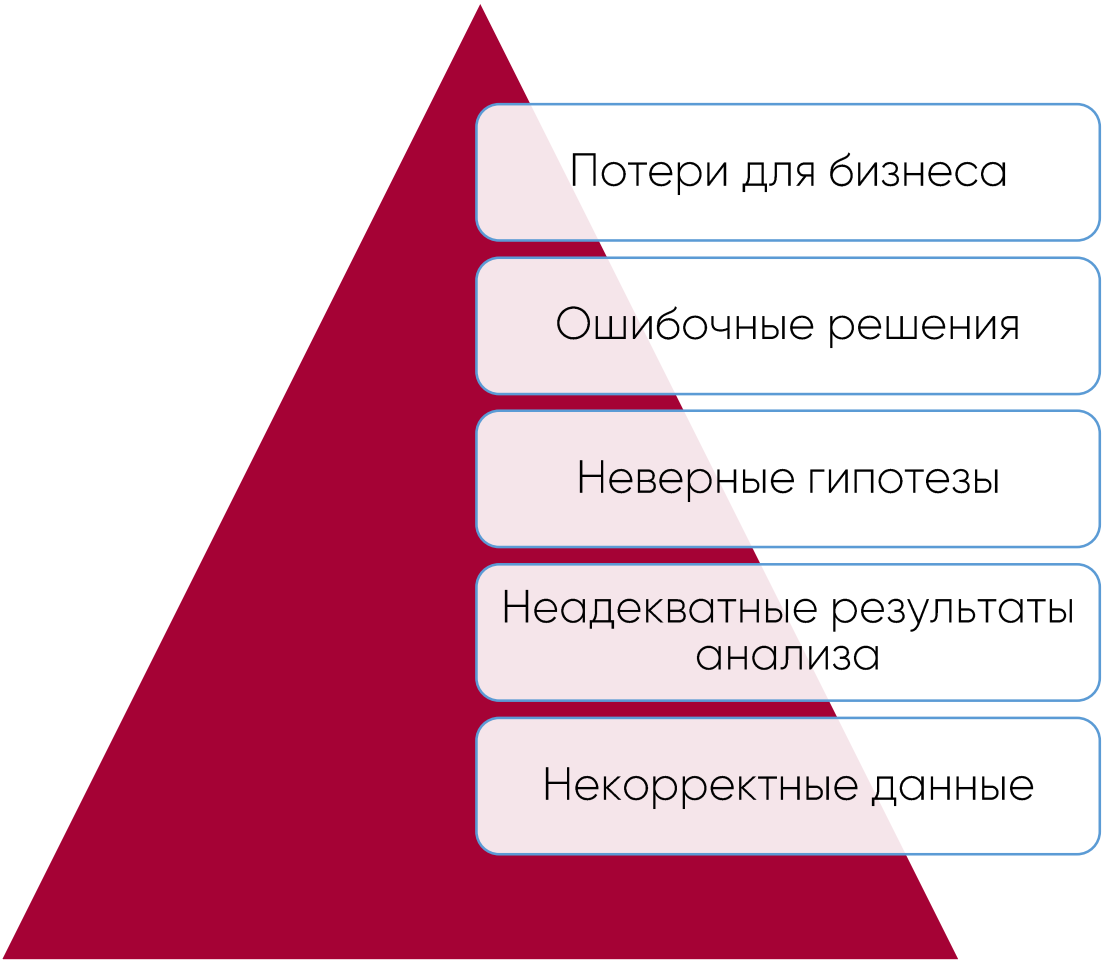
Комплекс организационной структуры, процессов, практик, подходов и участников, обеспечивающий выполнение требований к данным.

Инфраструктура управления данными.

Включает в себя основные программно-аппаратные комплексы, осуществляющие обработку данных, а также системы, имеющие критическую важность для обеспечения доступности и качества данных.

Качество данных: почему это важно?

Ошибки в данных могут привести к выработке ошибочной стратегии



Потери для бизнеса

Ошибочные решения

Неверные гипотезы

Неадекватные результаты
анализа

Некорректные данные

Как это может быть:

Ошибка в биржевых данных -> Ошибка моделей
-> Убыточная сделка

Ошибка в метеорологических данных -> Прогноз
хорошей погоды на выходных -> Пострадавшие
на природе отдыхающие

Ошибка в клиентских данных -> Неверная
атрибуция -> Финансирование терроризма

Ошибка в записи дня рождения тещи (свекрови)
-> Отсутствие поздравлений -> Испорченное
настроение

Элементы системы автоматизированного контроля КД



Ввод информации
(системы-источники)

Консолидация данных
(платформы и фабрики
данных)

Предоставление
данных потребителям
(специализированные
системы и витрины)

Элементы системы автоматизированного контроля КД



Элементы системы автоматизированного контроля КД



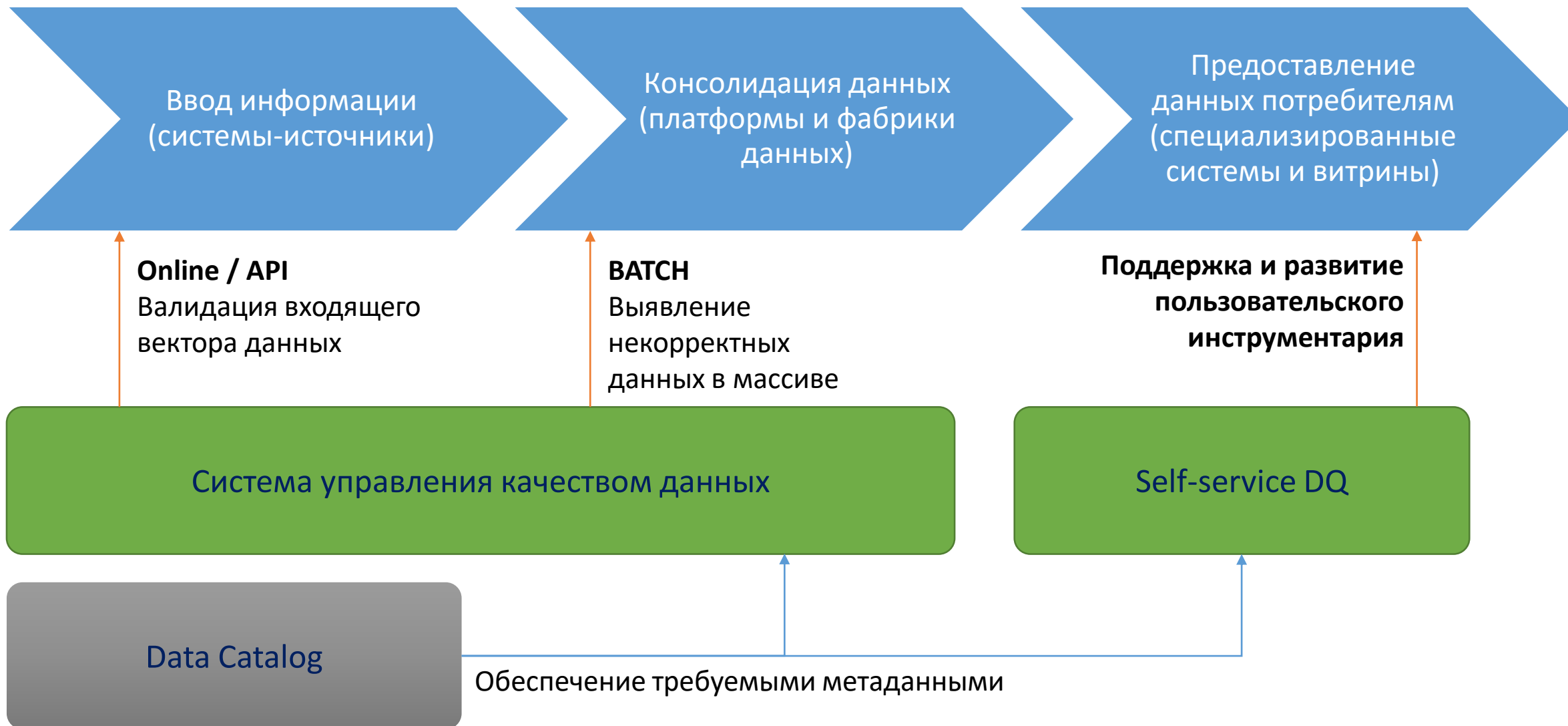
Проблематика стандартного подхода



Поиск альтернативных решений



Системы Офиса CDO в альтернативном подходе



Паттернализация пост-контролей: проблематика

На различных слоях платформ консолидации имеются данные, относящиеся к одному и тому же атрибуту корпоративной модели данных.

Пример:

Реквизиты клиента ФЛ
(источник CRM)

Реквизиты клиента ФЛ
(источник АБС)

Реквизиты клиента ФЛ
(источник CDI)

Реквизиты клиента ФЛ
(источник CDI)

Реквизиты клиента ФЛ
(витрина Риски)

Реквизиты клиента ФЛ
(витрина вкладчики)

Реквизиты клиента ФЛ
(витрина ...)

Физическая модель хранения разная на каждом источнике/слое, но правила контроля качества данных - одинаковые

Паттернализация пост-контролей: решение

Пост-контроли качества реализовать в терминах корпоративной модели данных, а реализацию для конкретной точки контроля генерировать автоматически на данных маппинга на физические модели (из каталога данных).

Основная информация	Области контроля	Запрос	Даты	Характеристика данных	Триггеры проверок	Периодичность
Ключ филиала:		Произвольный атрибут группировки результатов: ОСНОВНЫЕ СВЕДЕНИЯ.Статус клиента				
Запрос выражение 1:	ОСНОВНЫЕ СВЕДЕНИЯ.ФИО одной строкой					
Запрос выражение 2:	ОСНОВНЫЕ СВЕДЕНИЯ.Идентификатор в АБС					
Запрос выражение 3:	ОСНОВНЫЕ СВЕДЕНИЯ.Унифицированный идентификатор GUID					
Запрос проверки (WHERE): *	<input '[0-9!#\$%&\\'+="" =="" ?^`{}~]')="" \"основные="" and="" is="" not="" null"="" type="text" value="regexp_like ('\" одной="" основные="" сведения.фио="" строкой\"="" строкой\",=""/>					

Основная информация	Области контроля	Запрос	Даты	Характеристика данных	Триггеры проверок	Периодичность
Объект контроля	Область контроля	Входит в индикатор	Запрос			
ОСНОВНЫЕ СВЕДЕНИЯ	DQ_PREP	N	SELECT CAST(decode(party_rk, -1, party_global_rk, party_rk) AS string) AS key_id, CAST(CAST(t_source_system_id AS string) AS string) AS source_system_id, NULL AS l			

Контроль входящих данных систем первичного ввода информации: проблематика



Пост-контроли качества данных позволяют выявлять и устранять значительное количество ошибок в данных, но только блокирование ввода заведомо некорректной информации может этот поток ошибок остановить.

- В крупных организациях данные вносятся во множество систем, включая вендорские решения. Возникает большая управленческая нагрузка на ведение таких изменений.
- Фронтальные системы – это системы, связанные непосредственно с бизнесом, поэтому доработки, связанные с качеством данных имеют более низкий приоритет.
- Реализация контролей качества в системах нередко требует изменений в интеграции и совместных релизов.
- Требования к контролям качества данных достаточно часто изменяются.

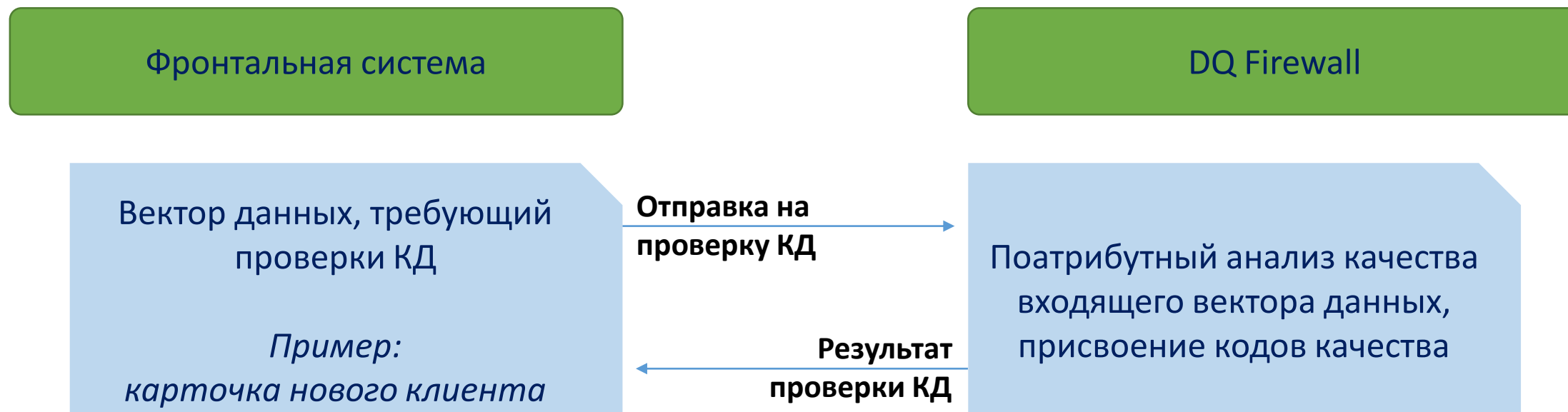
Контроль входящих данных систем первичного ввода информации: data quality firewall



Идея:

У нас есть библиотека пост-контролей в терминах корпоративной модели данных, которую мы развиваем и актуализируем.

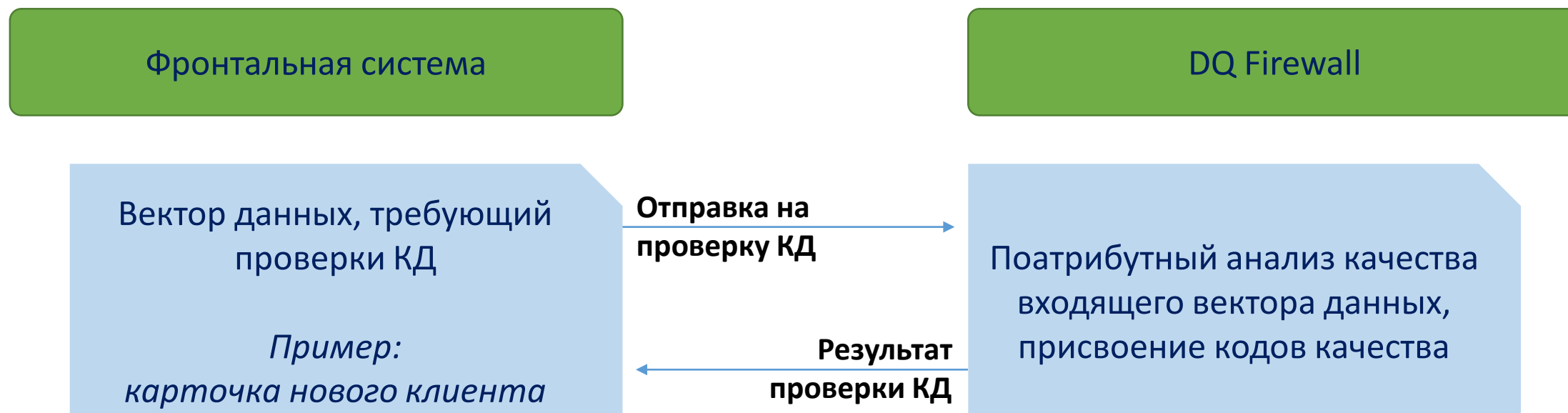
А если сделать сервис (API) для фронтальных систем по валидации данных в реальном времени по запросу? Например, как REST-сервис?



Data Quality Firewall : концепция



1. Синхронный сервис, получающий на вход вектор данных для валидации, возвращающий поатрибутный ответ в кодах качества.
2. Использует библиотеку контролей КД, разработанную для пост-контроля.
3. Для ускорения обслуживающие процессы (в т.ч. логирование) выведены в асинхронный режим и отделены от основного процесса.



Data Quality Firewall: пример работы сервиса

Вектор данных:

```

{
  "ГРАЖДАНСТВО.Страна": "РФ",
  "ОСНОВНЫЕ СВЕДЕНИЯ.ФИО одной строкой": "Иванов Иван Иванович",
  "ДУЛ.Паспорт РФ.Тип ДУЛ": "21",
  "ДУЛ.Паспорт РФ.Кем выдано": "111-111",
  "АДРЕС.Страна": "Российская Федерация",
  "АДРЕС.Почтовый индекс": "111111",
  "АДРЕС.Город (код ФИАС)": "г Москва",
  "ДУЛ.Паспорт РФ.Дата выдачи": "10.02.2003 0:00:00"
}

```

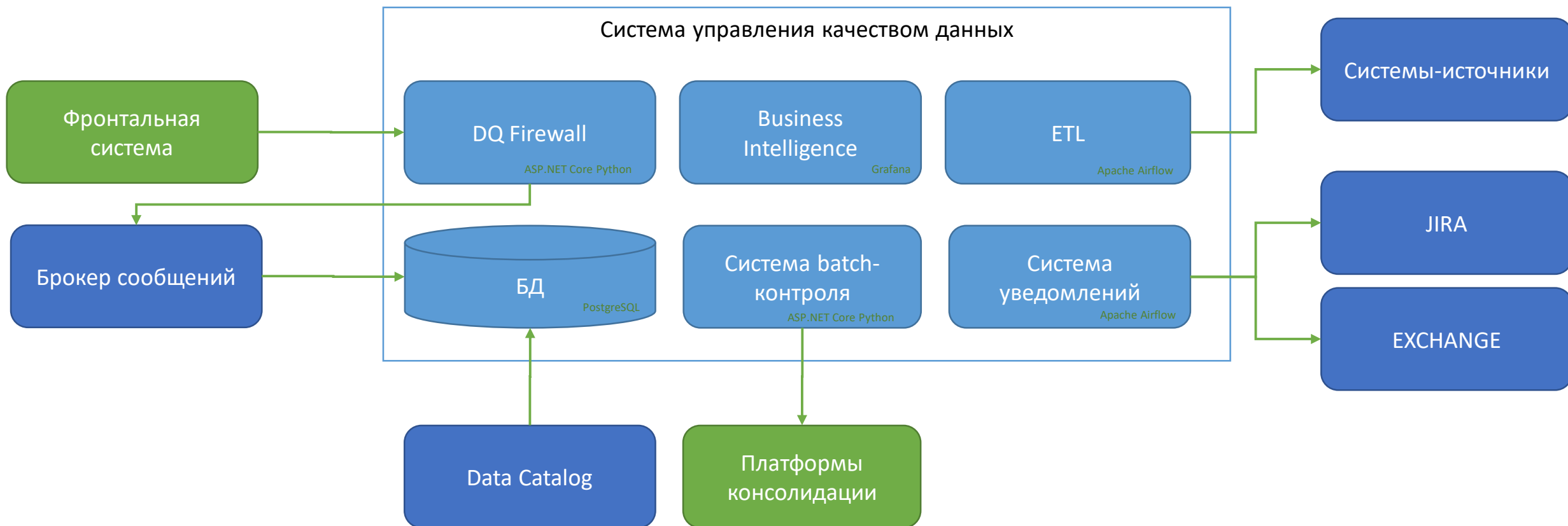
Результат:

SUCCESS

Проверить

Результат по проверкам			Результат по проверкам	Результат по атрибутам
ID проверки	Наименование	Запрос проверки (WHERE)	Атрибут	Результат
1088	ФЛ / ФИО / Содержит кириллицу и латиницу	"ОСНОВНЫЕ СВЕДЕНИЯ.ФИО одной строкой" regexp '[a-zA-Z]+' AND "ОСНОВНЫЕ СВЕДЕНИЯ.ФИО одной строкой" regexp	АДРЕС.Почтовый индекс	SUCCESS
10015	ФЛ / Адрес / Индекс / Состоит не из 6 цифр для РФ	"АДРЕС.Страна" = 'Российская Федерация' and trim("АДРЕС.Почтовый индекс") not REGEXP '^([0-9]){6}\$'	ДУЛ.Паспорт РФ.Кем выдано	SUCCESS
1087	ФЛ / ФИО / Пустое	"ОСНОВНЫЕ СВЕДЕНИЯ.ФИО одной строкой" is null	ОСНОВНЫЕ СВЕДЕНИЯ.ФИО одной строкой	SUCCESS
1165	ФЛ / Паспорт РФ / Орган выдачи менее 5 символов 2	"ДУЛ.Паспорт РФ.Тип ДУЛ" = '21' AND LENGTH("ДУЛ.Паспорт РФ.Кем выдано") < 5	ГРАЖДАНСТВО.Страна	SUCCESS
10009	ФЛ / Паспорт РФ / Дата выдачи / Пустая	"ДУЛ.Паспорт РФ.Тип Дул" = '21' AND "ДУЛ.Паспорт РФ.Дата выдачи" IS NULL	АДРЕС.Город (код ФИАС)	SUCCESS
1085	ФЛ / Гражданство / Пусто или Unknown	"ГРАЖДАНСТВО.Страна" IS NULL OR "ГРАЖДАНСТВО.Страна" = 'unknown'	ДУЛ.Паспорт РФ.Дата выдачи	SUCCESS
10034	ФЛ / Адрес / Город / Начинается с дефиса или пробела9	regexp_like("АДРЕС.Город (код ФИАС)", '^[-]')		

Data Quality Firewall: архитектура



Пост-контроль в сервисной модели

Офис CDO

- Владеет и развивает библиотеку контролей качества данных в терминах корпоративной модели данных.
- По мере необходимости подключает новые объекты контроля КД и предоставляет на регулярной основе результаты контролей качества данных.
- Обработывает выявленные инциденты качества данных

Потребитель сервиса

- Использует результаты контроля качества данных для своих целей
- Учитывает факты возникновения инцидентов качества данных в своей работе

DQFirewall в сервисной модели

Офис CDO

- Владеет и развивает библиотеку контролей качества данных в терминах корпоративной модели данных.
- Развивает и сопровождает программный сервис (API) по валидации качества данных.
- Осуществляет валидацию входящего вектора данных и возвращает поатрибутный ответ о соответствии/несоответствии данных требованиям по качеству.

Потребитель сервиса

- Осуществляет подключение к программному сервису (API) своего бизнес-приложения.
- Программно направляет запрос на валидацию данных.
- Обрабатывает и выводит на экранные формы результаты работы сервиса.

Сервисная модель : сложности подхода

1. *Основные потребители сервиса (DQ Firewall API) – системы Mission Critical. Должен быть обеспечен аналогичный уровень отказоустойчивости.*
2. *Должно быть низкое время отклика такого сервиса.*
3. *Для batch-контроля должно быть обеспечено высокое качество маппинга в каталоге данных корпоративной модели данных на физические модели точек контроля.*
4. *Необходима готовность команды к большому объему интеграционных задач.*

Сервисная модель : выгоды и плюсы

1. *Радикальное сокращение time-to-market на изменения в контролях качества данных (достаточно внести изменения через интерфейс системы).*
2. *Снижение управленческих затрат за счет единой точки управления изменениями в контролях качества данных.*
3. *Единая методология контроля КД на всех узлах контроля. Используется сквозная библиотека контролей. Исключена рассинхронизация версий контролей – при изменении происходит регенерация исполняемого кода и обновление кэша для API.*
4. *Появление нового узла контроля – типовая задача.*
5. *Возможность внедрения дополнительных инструментов контроля КД (в том числе интеграции с ГИС) прозрачно для потребителей сервисов.*



Спасибо за
внимание!

gazprombank.ru

