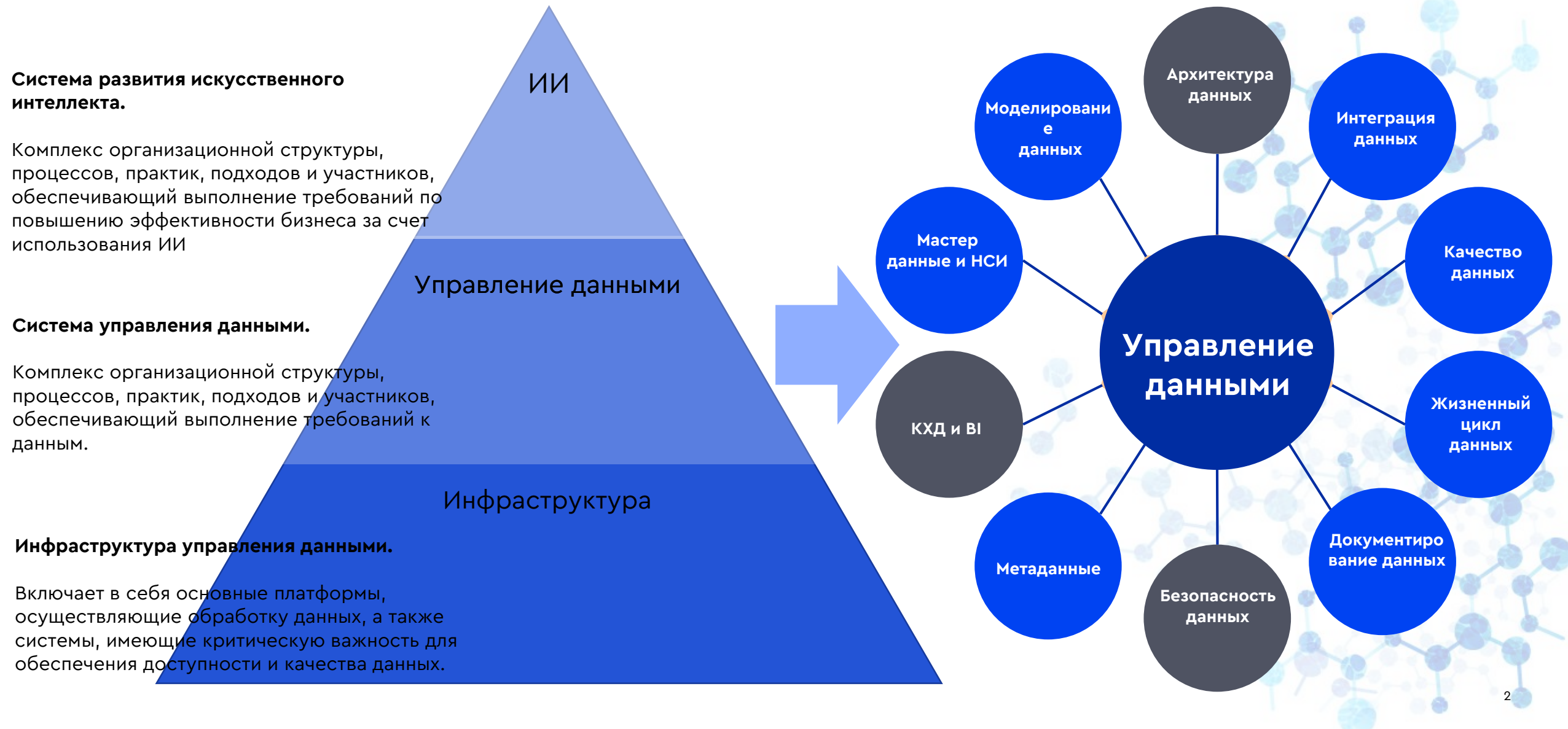




Технологический суверенитет: опыт построения импортонезависимой платформы управления данными

Бондаренко Алексей Николаевич
Вице-президент – Начальник Департамента управления данными
Газпромбанк





Ключевые задачи управления данными

Бизнес-драйверы

Сокращение
трудозатрат в
процессах
обработки данных

Повышение качества
клиентского опыта

Повышение точности
оценки рисков

Увеличение продаж

Повышение качества данных

Повышение доступности данных

Повышение информированности о данных

Сокращение трудозатрат на обработку данных

Развитие AI и ML

Переход на новую
инфраструктуру
данных

Развитие
микросервисной
архитектуры

Импортозамещение

Технологические драйверы

Что такое Платформа управления данными?

Это комплекс инструментов для решения вышеперечисленных задач:

- **Повышение качества данных**
- **Повышение доступности данных**
- **Повышение информированности о данных**
- **Сокращение трудозатрат на обработку данных**

В нашем случае вне контура Платформы управления данными остаются:

- **Система хранения и обработки данных (DWH, DataLake, etc.)**
- **Инструменты обеспечения безопасности данных**

Строить или покупать

Чем обусловлен выбор в пользу собственной разработки:

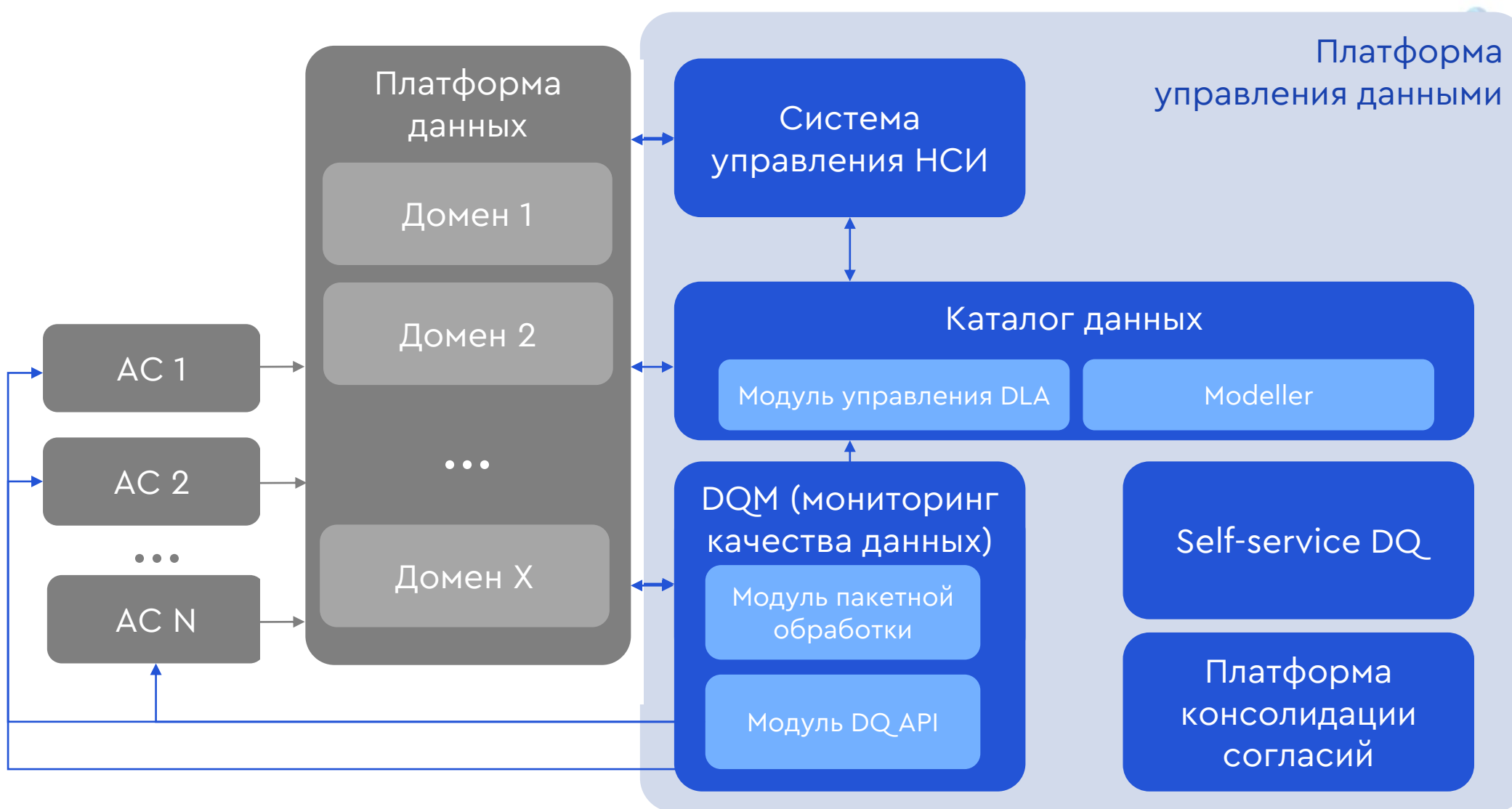
Собственная разработка:

- ☒ Точная реализация требований
- ☒ Лучший T2M доработок
- ☒ Гибкость в выборе техстека
- ☒ Лучше TCO на длинном сроке
- ☒ Выше первоначальные инвестиции
- ☒ Необходимо разработать процессы и методологию
- ☒ Необходима собственная техническая команда

Покупка готового решения:

- ☒ Ниже первоначальные инвестиции
- ☒ В решении уже реализованы процессы и методология
- ☒ Не нужна собственная техническая команда
- ☒ Необходимо подстраивать свои требования под возможности системы
- ☒ T2M увеличивается на продолжительность задач по контрактам и оплате
- ☒ Возможные сложности при интеграции систем от различных вендоров
- ☒ На длинном сроке, как правило, TCO имеет свойство расти

Платформа управления данными



Драйверы



Создание базы знаний о данных компании



Экономия трудозатрат аналитиков



Закрепление ответственности за данные



Экономия трудозатрат ИТ



Импортозамещение

Задачи



Централизация

Центральное хранилище метаданных, способное собирать метаданные из различных СИ. Каталогизация и описание отчетов (BI). Единый реестр бизнес-терминов (гlossарий).



Технический Lineage

Сбор метаданных в рамках открытого стандарта OpenLineage в части процессов сбора и обработки данных. Изучение показателей производительности и деталей выполнения процессов



Бизнес-Lineage

Поддержка и визуализация полной родословной для всех данных. Отслеживание пути данных по таблицам, конвейерам и BI отчетам.

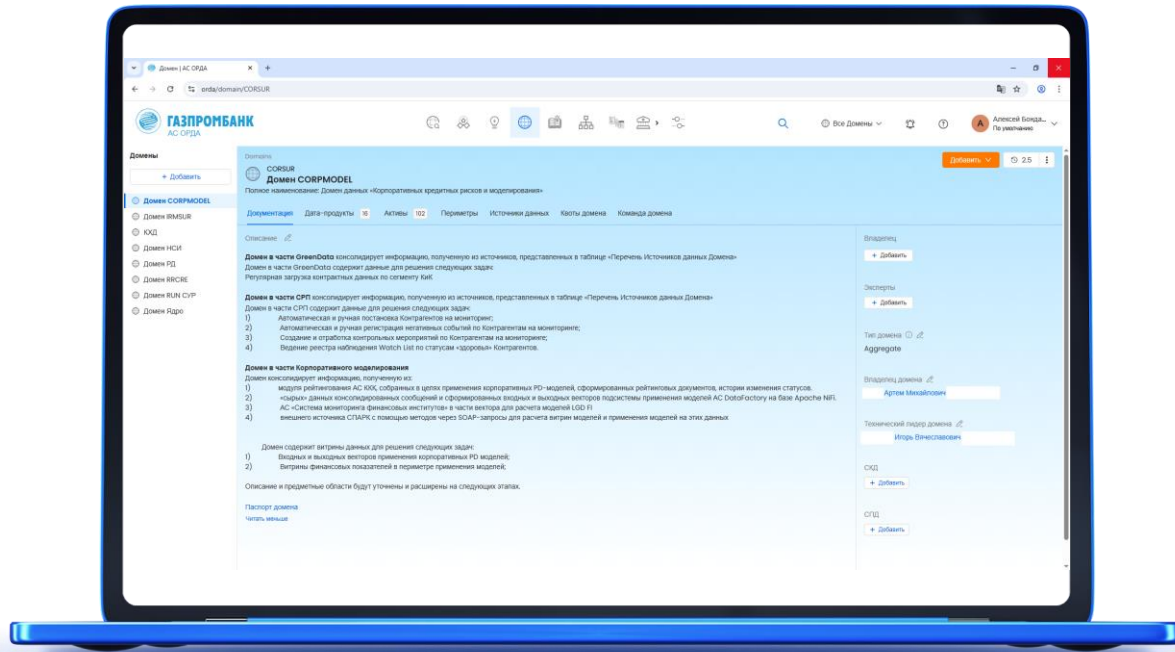


Совместная среда работы

Сотрудничество между разработчиками и потребителями. Масштабируемый и полнотекстовый поиск. Лента изменений активности, позволяет просматривать сводку событий изменения в данных

Имплементация доменного подхода

Переход на новую концепцию создания хранилища (DataMesh) потребовал доработки платформы управления данными для выполнения следующих задач:



Разделение данных (объектов) на Домены



Реализация классификатора Дата-продуктов



Формирование ролевой модели доменных команд



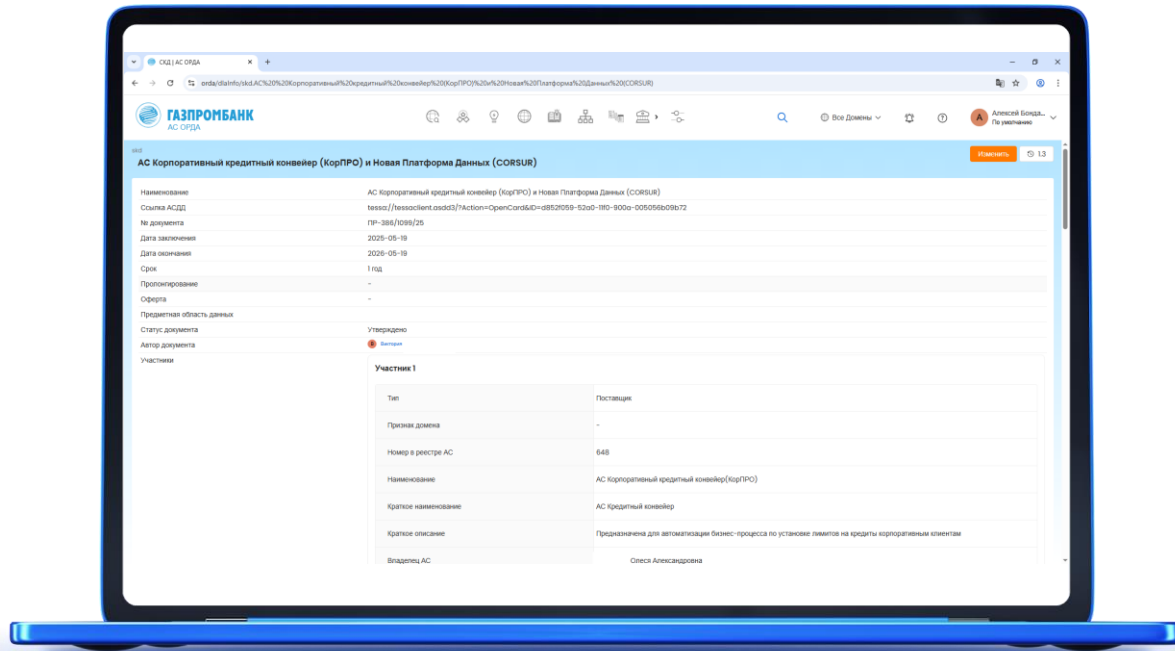
Обеспечение для каждой доменной команды собственной среды разработки объектов и контролей



Обеспечение разграничения доступа команд к централизованным сервисам

Модуль управления DLA

Инструмент разработки и заключения соглашений о поставке и качестве данных



Создание соглашений о поставке и качестве данных в режиме конструктора



Возможность использования при проектировании DLA уже загруженных в Каталог данных объектов, информацию о подразделениях и сотрудниках

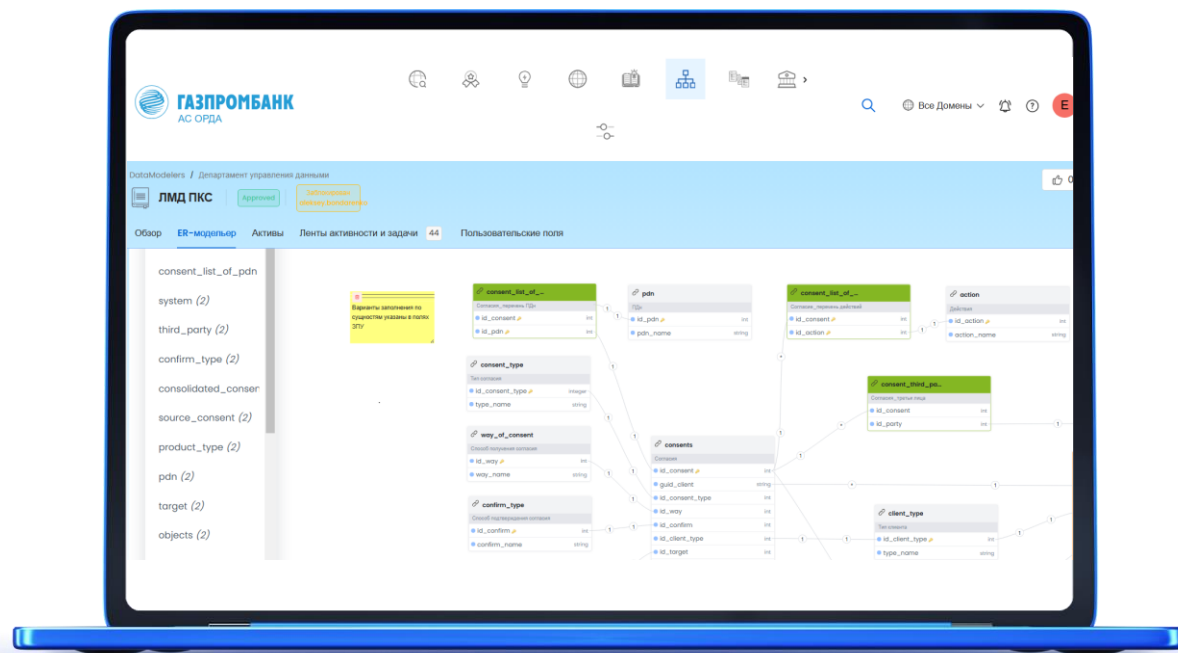


Наличие как внутреннего workflow согласования, так и возможности передачи на согласование в единую СЭД



Интеграция с DQM для автоматизированного контроля исполнения DLA

Импортонезависимый инструмент проектирования баз данных



Инструмент создания логических и физических моделей данных



Возможность использования при проектировании уже загруженных в Каталог данных объектов и создания новых (проектных)



Возможность проектирования потоков данных (s2t) на основе метаданных как существующих, так и проектных

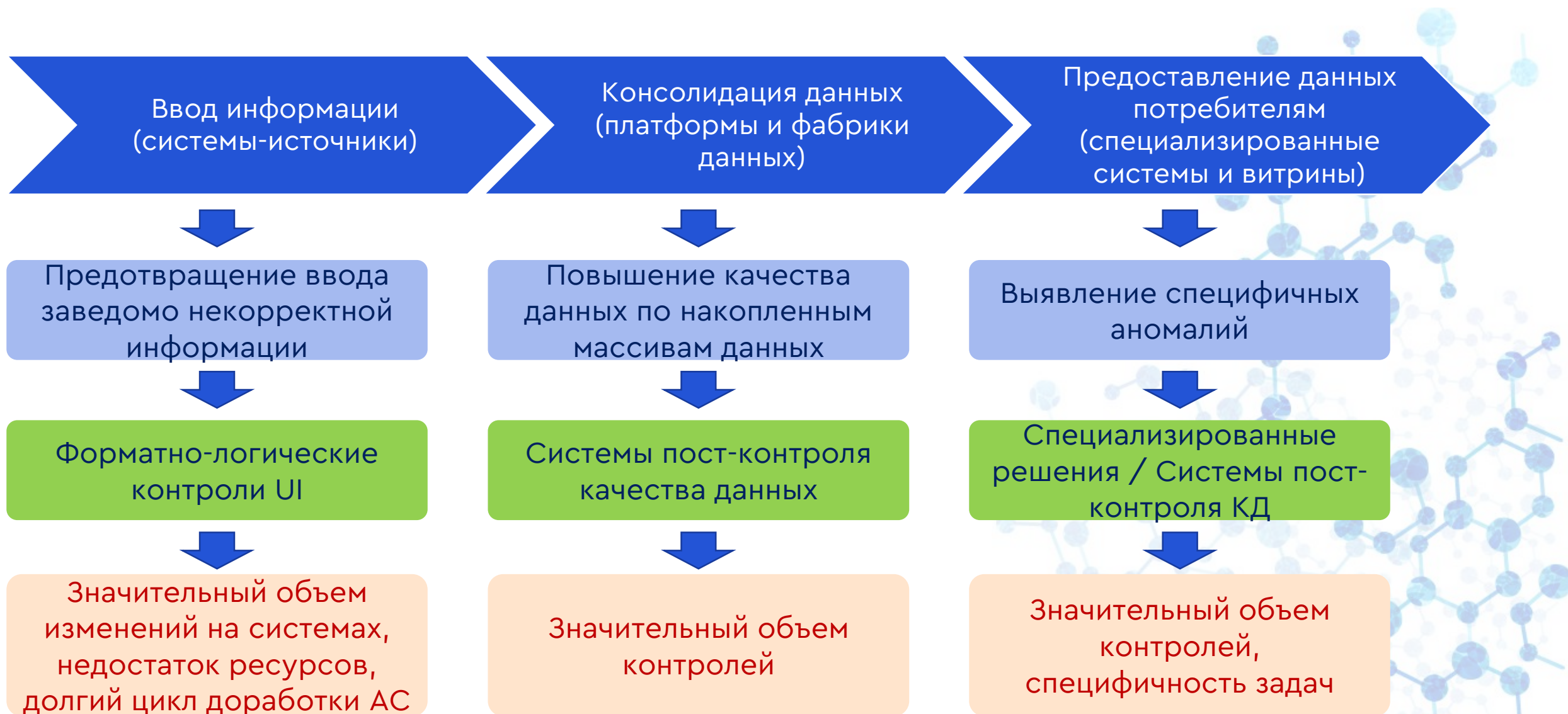


Возможность экспорта данных по результатам проектирования

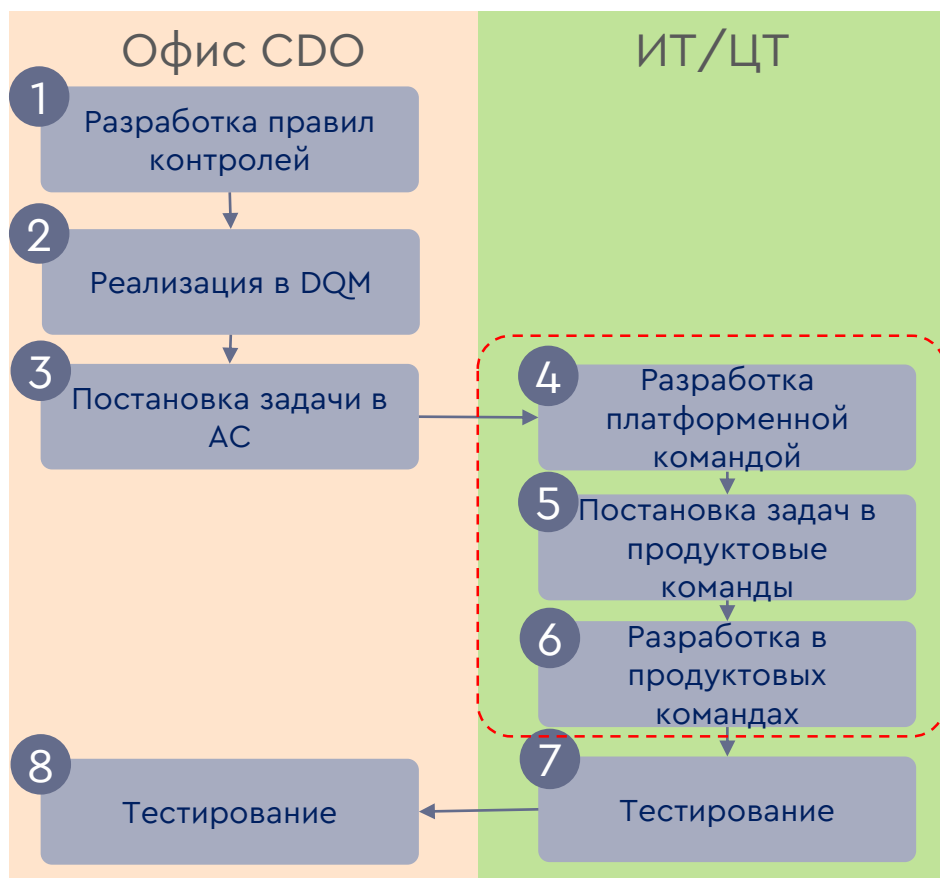


Генерация кода и работа в режиме API

Три уровня контроля качества данных

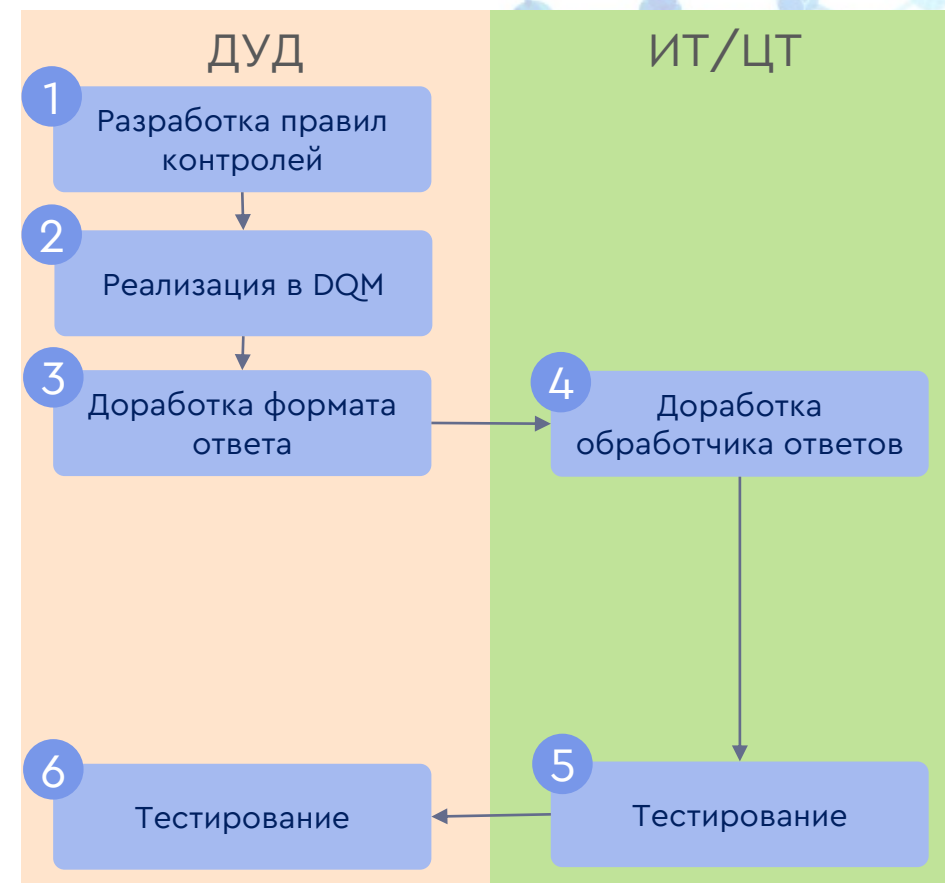


Старый подход – локальная реализация ФЛК

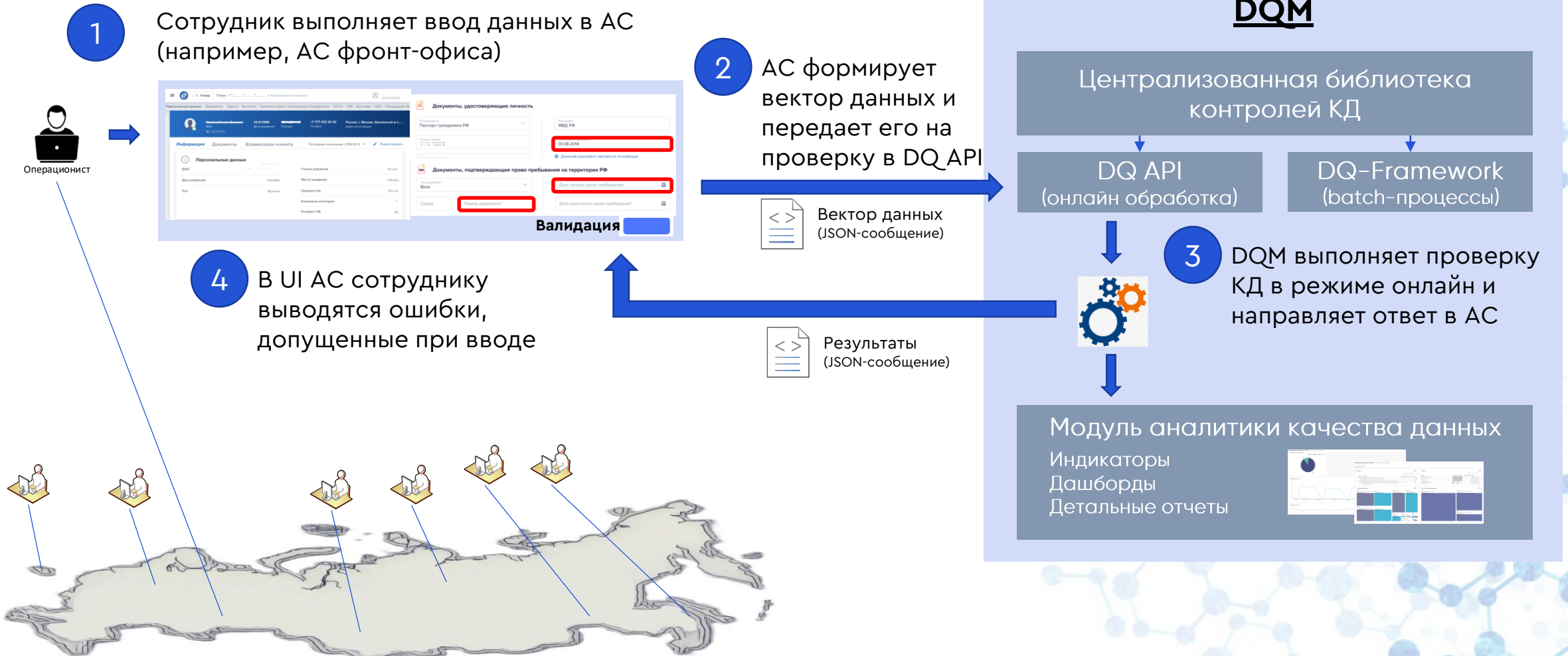


Исключается значительная нагрузка на команды АС.

Новый подход – DQ API

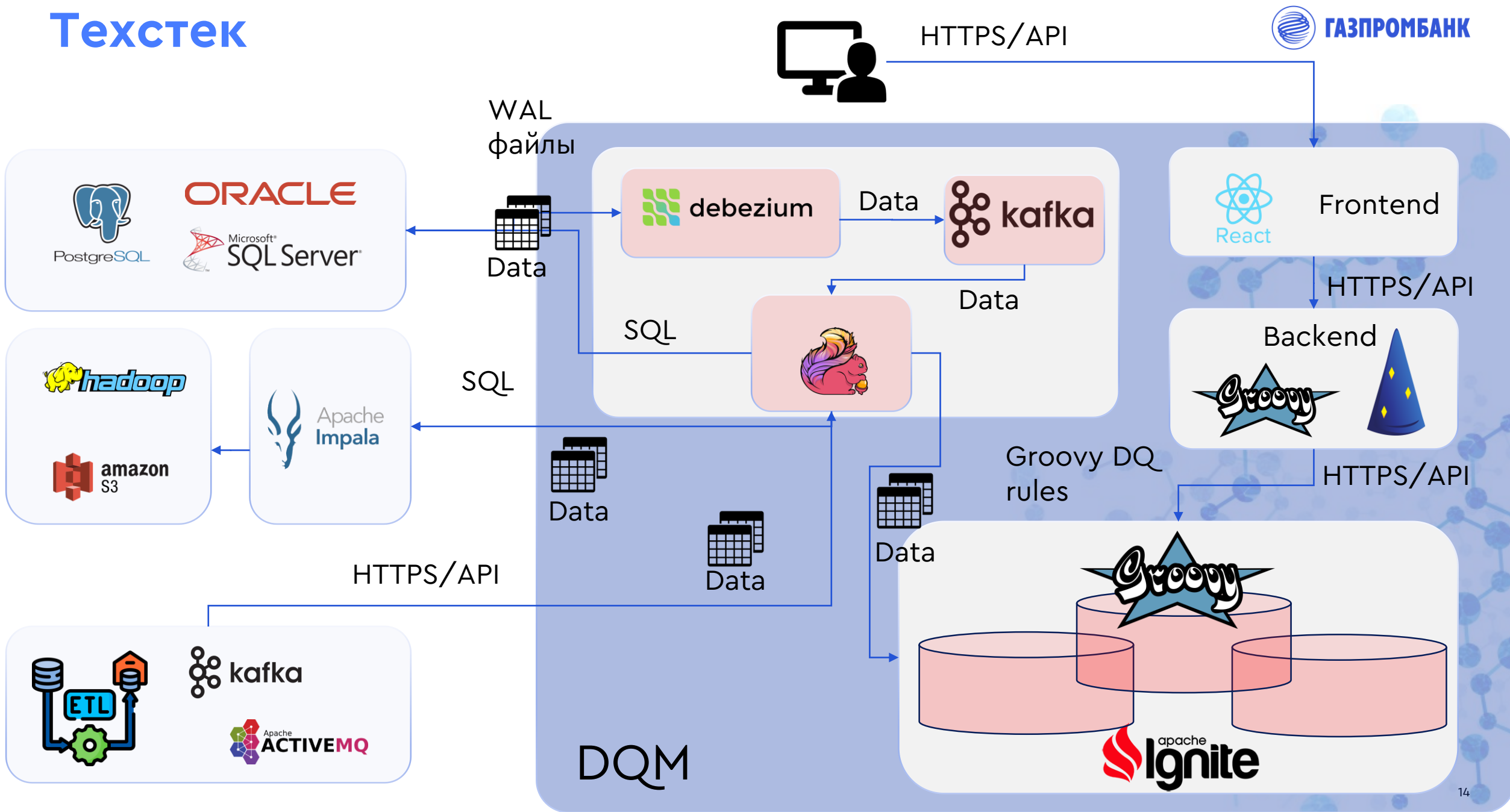


DQ API – как это работает?



Плюсы подхода:

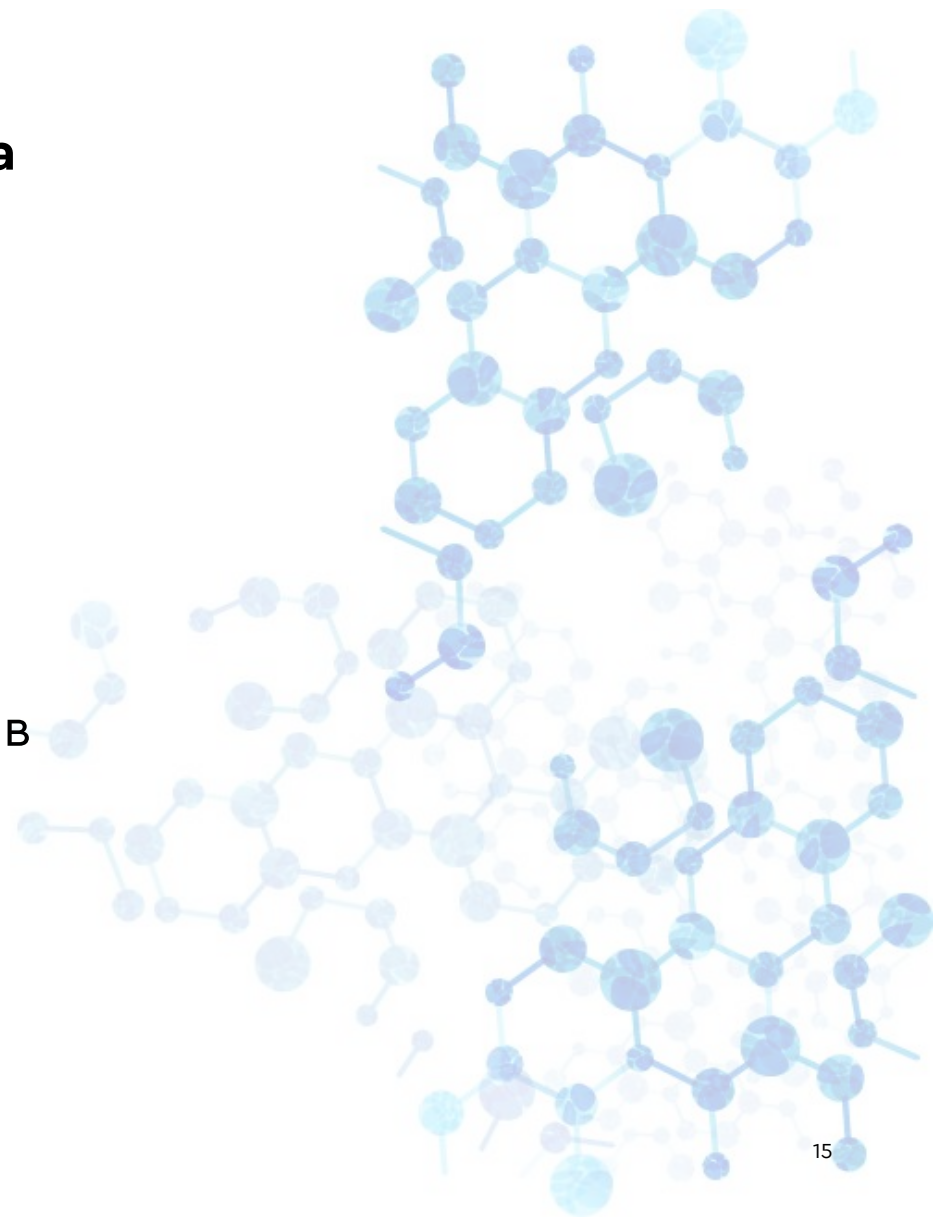
- Использование централизованных правил контроля
- Отсутствие необходимости поддержания полного и актуального набора контролей КД на стороне АС (Потребителя сервиса)
- Прозрачность, накопление статистики, построение аналитики о работе операторов на вводе данных в АС, в том числе с динамикой изменения (улучшения)
- Обеспечение качества данных в АС на новом потоке на уровне 99%++



В рамках программы цифровой трансформации в Банке была успешно внедрена импортонезависимая платформа управления данными, решающая задачи:

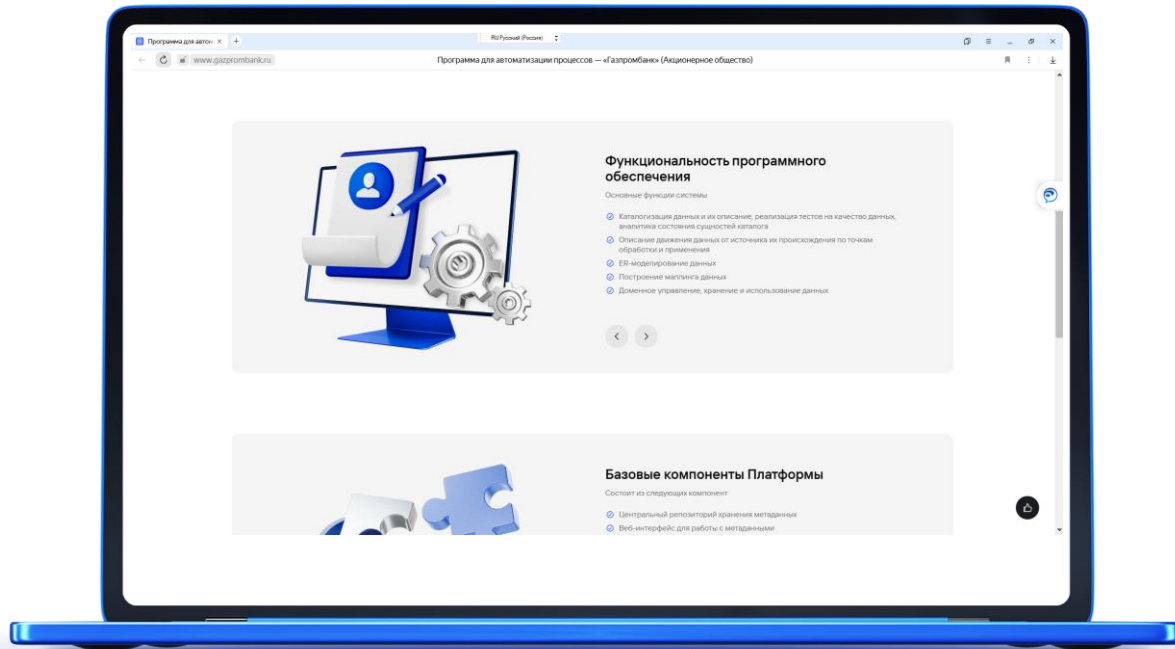
- Повышения качества данных
- Повышения доступности данных
- Повышения информированности о данных
- Сокращения трудозатрат на обработку данных

При внедрении приоритет отдавался широкому использованию зарекомендовавших себя open-source компонент (OpenMetadata, Debezium, Ignite, Groovy и т.д.) в сочетании с собственной разработкой функционала, в первую очередь с использованием языка Java.



Информационные материалы

<https://www.gazprombank.ru/special/a/automation-process/>
<https://www.gazprombank.ru/special/a/verification-system/>



Информация о системе



Руководство по эксплуатации



Руководство администратора
(инструкция по установке)