



Как мы сделали одну большую песочницу для всех аналитиков



Спикеры:
Алексей Кошевой, Кристина Проскурина

Алексей Кошевой, руководитель отдела развития витрин данных РСХБ-Интех

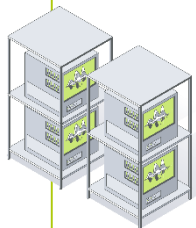
Более 10 лет опыта в ИТ — в направлении, связанном с DWH. Занимал разные позиции от аналитика и разработчика до руководителя DWH. Сфокусирован на архитектуре данных в хранилище, качестве, консистентности предоставляемых данных и простоте их анализа



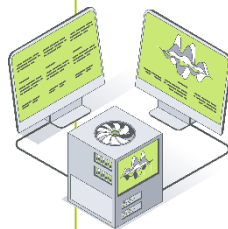
Кристина Проскурина, управляющий Директор РСХБ-Интех

Более 10 лет опыта в ИТ. В крупном интеграторе руководила направлением Big Data. На данный момент — 2 года в РСХБ руководит управлением по бизнес-анализу данных и проектами в части Корпоративного хранилища данных и Data Research Platform

КАКИЕ ПРОБЛЕМЫ РЕШАЛИ



Разрозненные по разным
песочницам единообразные
данные



Сопровождение
нескольких систем



Одни и те же данные
хранились на разных серверах
(место, железо)



Отсутствие единой системы
управления данными

С ЧЕГО МЫ НАЧИНАЛИ

ГЛАВНАЯ КНИГА



ДФИП – 9 000 таблиц

ORACLE

ДКБ – 2 000 таблиц

ORACLE

ЦКРР – 200 таблиц



Риски – 5 000 таблиц

ORACLE

ДО – 2 000 таблиц

ORACLE

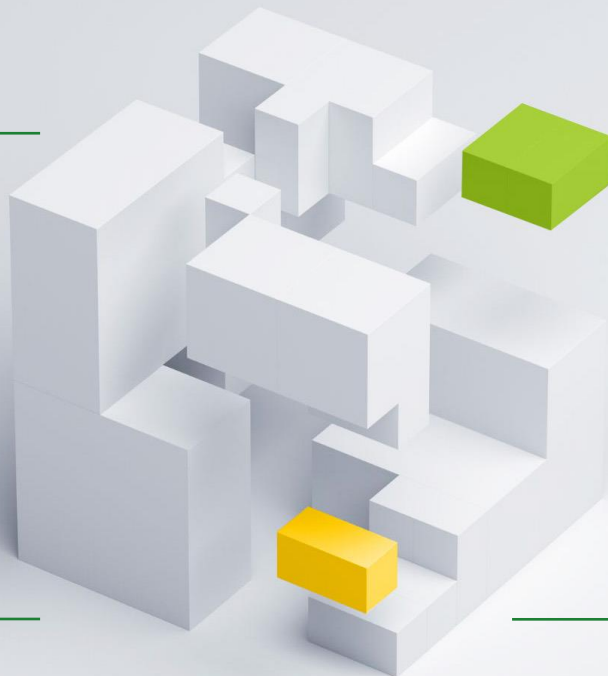
АСРМ – 2 000 таблиц

ORACLE

ДВК – много схем и таблиц



Розница – 2 000 таблиц



С ЧЕГО МЫ НАЧИНАЛИ



ДФИП — 9 000 таблиц

ИКРР — 200 таблиц

Риски — 5 000 таблиц

ДКБ — 2 000 таблиц

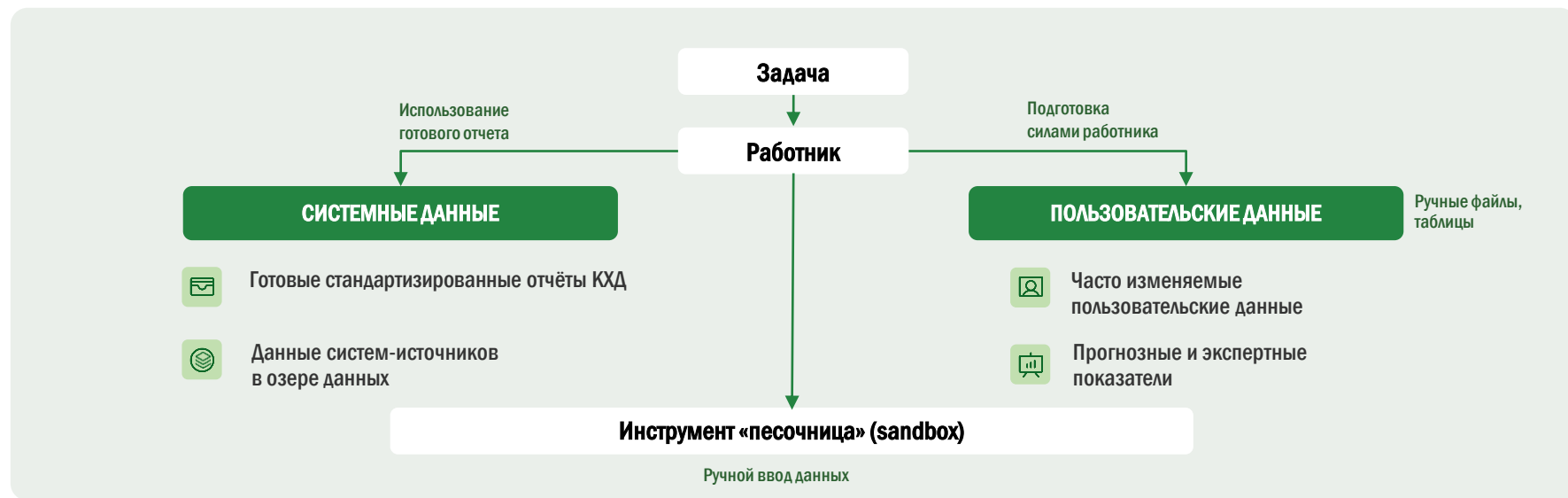
АСРМ — 2 000 таблиц

ДВК — много схем и таблиц

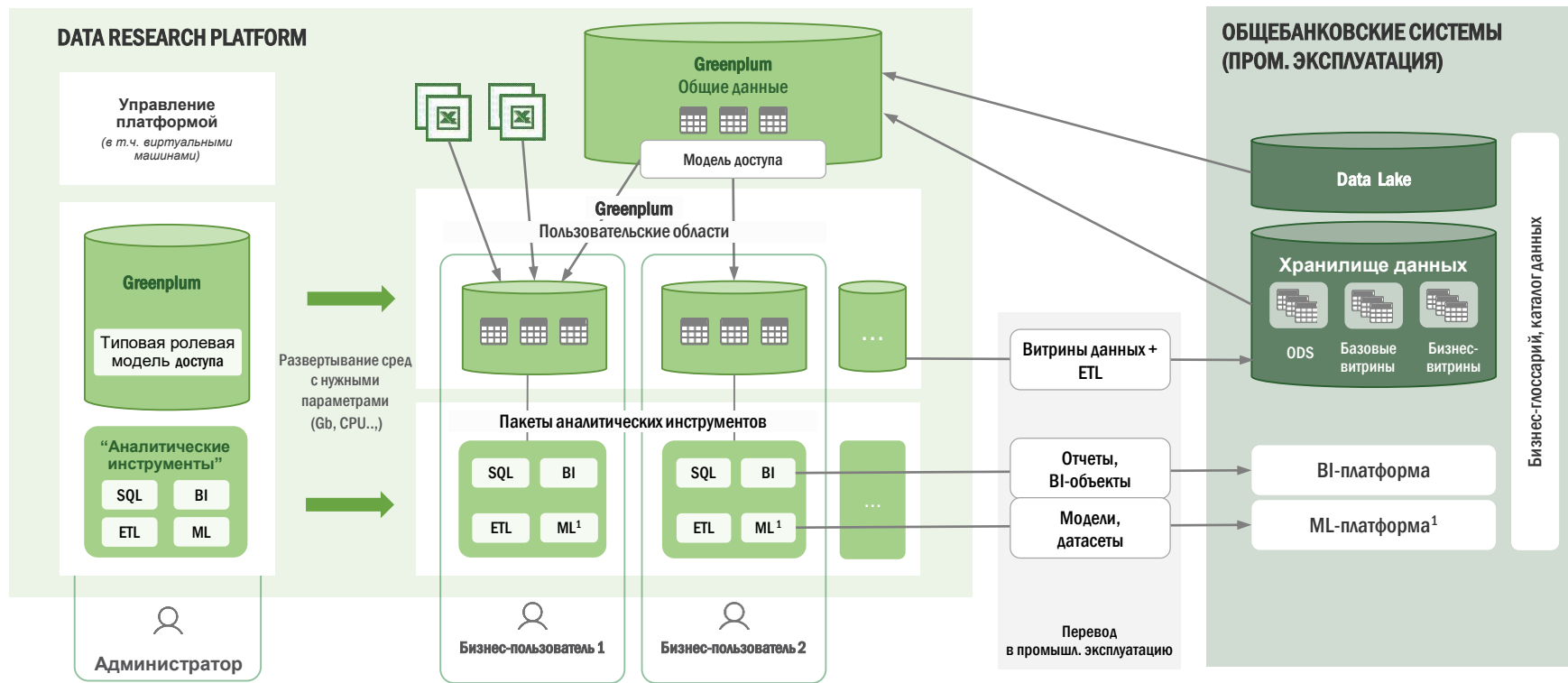
- Единый кластер GreenPlum Arenadata
- ETL Airflow, PXF
- Собственный движок
- Пространство - 100 ТБ
- 300 пользователей
- Миграция архивов и полное воспроизведение процедур с Oracle и MS SQL
- Сжатые сроки

ОДНИМ ИЗ ЦЕЛЕВЫХ ИНСТРУМЕНТОВ АНАЛИТИЧЕСКОГО КОНТУРА БАНКА ЯВЛЯЕТСЯ ИНСТРУМЕНТ ТИПА «ПЕСОЧНИЦА»

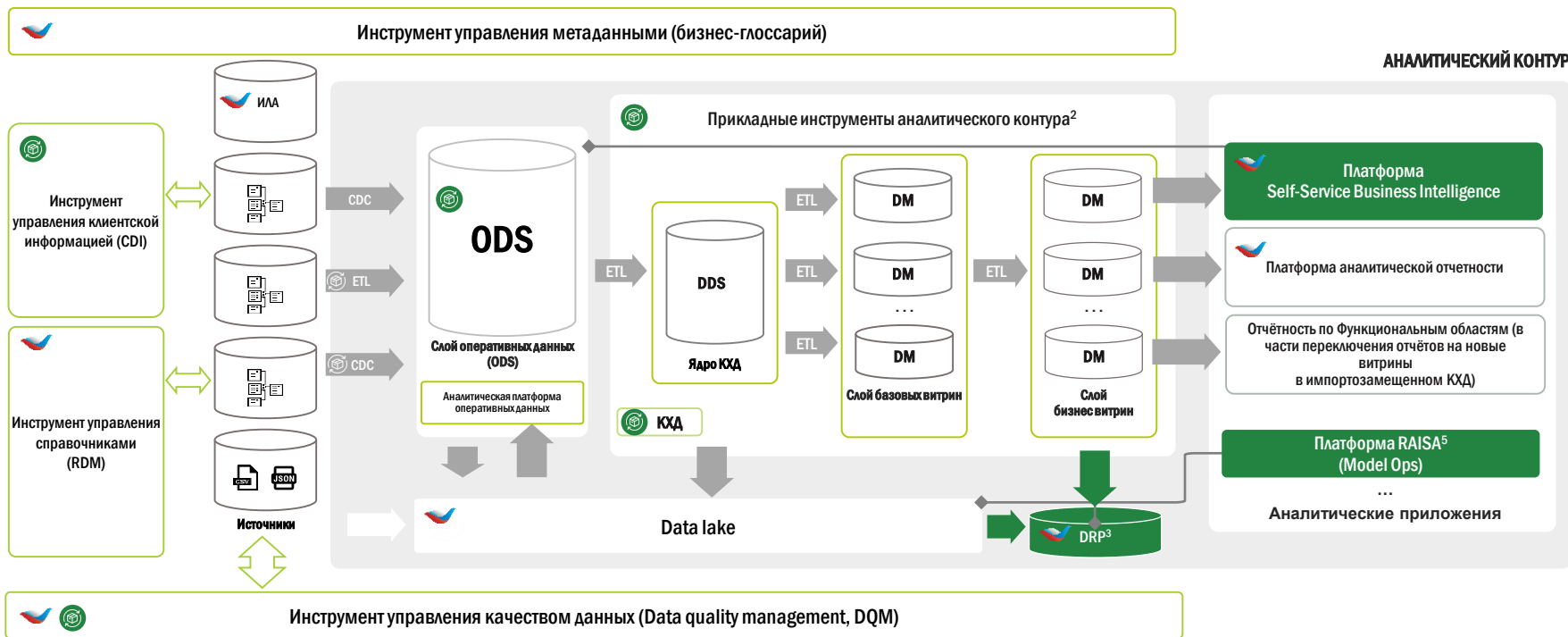
В зависимости от задачи, работник может воспользоваться системными данными (готовые данные аналитического контура), а может самостоятельно сформировать данные (загрузка локальных файлов, подготовка дополнительных расчетных таблиц на основе предзагруженных данных, загрузка внешних данных через локальные файлы).



ВЕРХНЕУРОВНЕВАЯ СХЕМА РЕАЛИЗАЦИИ DRP, ОТВЕЧАЮЩАЯ ЗАЯВЛЕННЫМ ПОТРЕБНОСТЯМ



ВЛИЯНИЕ ПРОЕКТА НА ЦЕЛЕВОЙ АНАЛИТИЧЕСКИЙ КОНТУР БАНКА



ПРОБЛЕМЫ

1. Отсутствовала документация по локальным песочницам;
2. За время работы сменилось несколько команд;
3. Большое количество объектов для миграции;
4. Желание бизнес пользователей переносить «как есть»;
5. Желание бизнеса делать все самостоятельно.



РЕШЕНИЯ

1. Разбили проект на функциональные области, которые сегментировали на стримы;
2. Сформировали бизнес требования к каждой из областей;
3. Сформулировали понятия «системные и пользовательские данные», для системных источников выделили команду, подготовили СПИР;
4. Для пользовательских данных подготовили инструмент.

ПРОБЛЕМЫ ПОСТАНОВКИ ТРЕБОВАНИЙ

Нам нужно все зеркало ЦХД (Хранилище данных)



1

Хорошо, но это 10000 таблиц.
Вы точно используете все 10000?



Хорошо, давайте
3 000 таблиц



2

Без проблем,
Но нужен список



Хорошо, нам достаточно
1 000 таблиц



3

Без проблем, но нужен список. Может
выделим бизнес область ?



... 50 встреч спустя...

Нам подойдет базовый слой КХД и 30
таблиц из ЦХД (список предоставим)



4

Отлично, пойдём делать



ПРОБЛЕМЫ

1. Выбор интеграции с источниками: выбирали между python + nfs, Spark и PXF
2. Разработка движка загрузки и оркестрации данных
3. Коммуникация с архитекторами
4. Быстрый рост функциональности системы: беклог насчитывал в моменте свыше 30 задач на движок



РЕШЕНИЯ

1. Протестировали каждый вариант, остановились на PXF + python
2. Переписывали движок: сначала он был монолитным, потом разбили на отдельные функции
3. Отказались от \$ в названиях
4. Приоритизировали задачи

ПРОБЛЕМЫ

1. Ролевая модель доступа к данным на ГП
2. Назначение владельцев данных
3. Ролевая модель доступа к дагам в AirFlow
4. Неконтролируемая загрузка данных



РЕШЕНИЯ

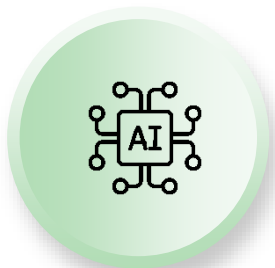
1. Разработали ролевую модель доступа к данным: общепанковские (есть доступ у всех), и данные, которые появляются в рамках работы каждого из бизнес направлений
2. Решили административно
3. Коллеги из ИИ помогли с решением
4. Квотирование пространств для бизнеса по объему

НА КАКОЙ СТАДИИ СЕЙЧАС?



Основные работы по
движку завершили

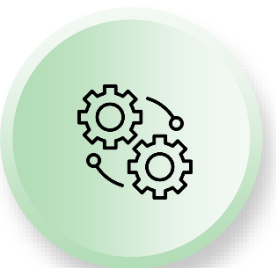
Используем механизм загрузки
данных используя ИИ
инструмент (Раиса)



Переносим песочницы, проводим
еженедельные митинги с каждой
командой из бизнеса

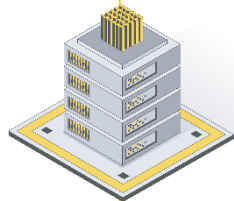


Ежедневно загружаем
2500+ объектов из разных
систем

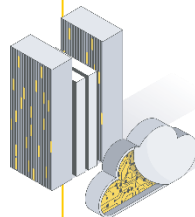


Организуем новые
интеграции

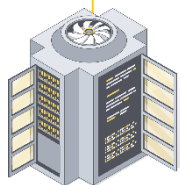
ЧЕГО ДОСТИГЛИ В РЕЗУЛЬТАТЕ ПРОЕКТА



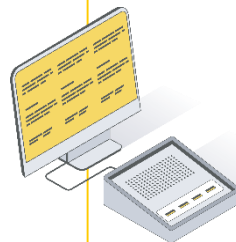
Единое место исследования и анализа данных, а также подготовка ad-hoc запросов



Одна система, которую необходимо сопровождать и развивать

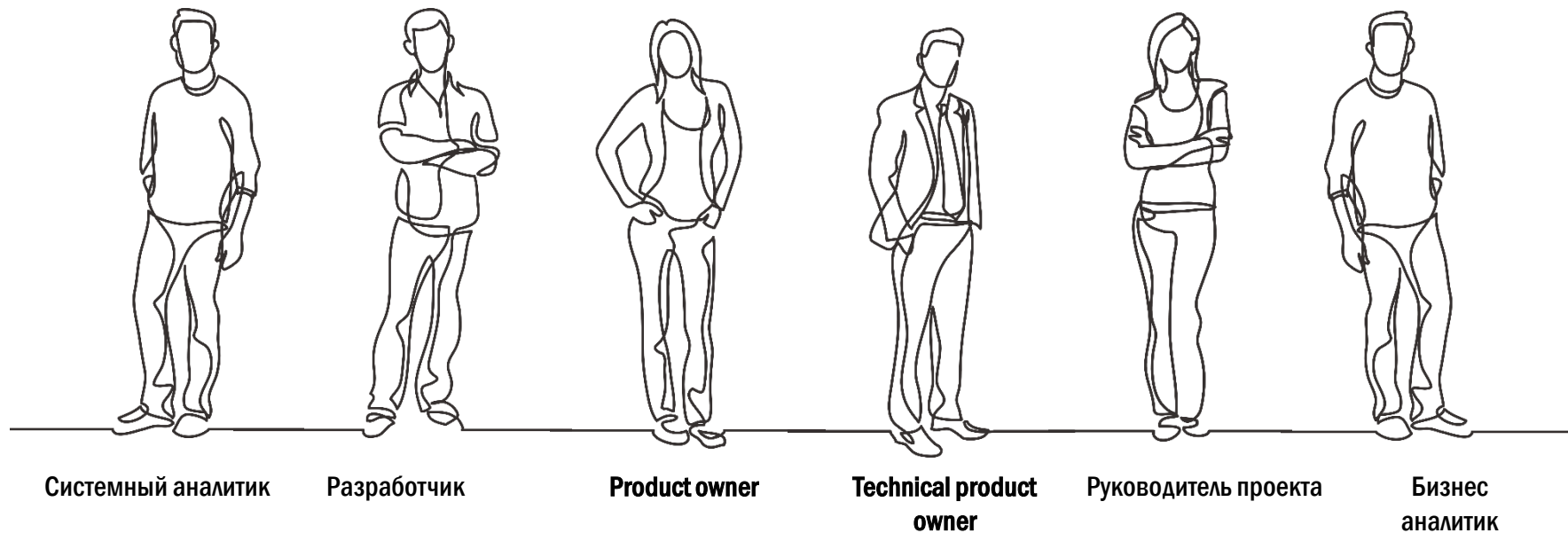


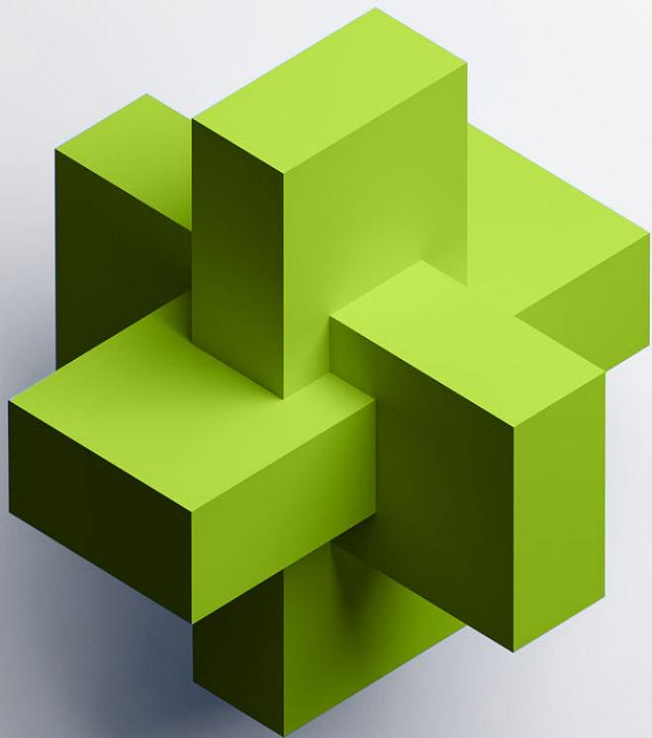
Единая система по управлению данными как с точки зрения IT, так и с точки зрения бизнеса



Описание всех объектов, загруженных в систему, актуализация на confluence и описание в бизнес глоссарии

СОСТАВ КОМАНДЫ ПРОЕКТА





**Спасибо
за внимание!**



ВАШИ ВОПРОСЫ