

Применение искусственного интеллекта в обработке административных данных



Владимир Некрасов, ООО «Контур Компонентс»

Предпосылки производства статистики из административных данных

- Взрывной рост данных в госуправлении
- Развитие информационных технологий и коммуникаций
- Новые возможности для статистики
 - Доступ к новым источникам
 - Повышение точности
 - Повышение оперативности

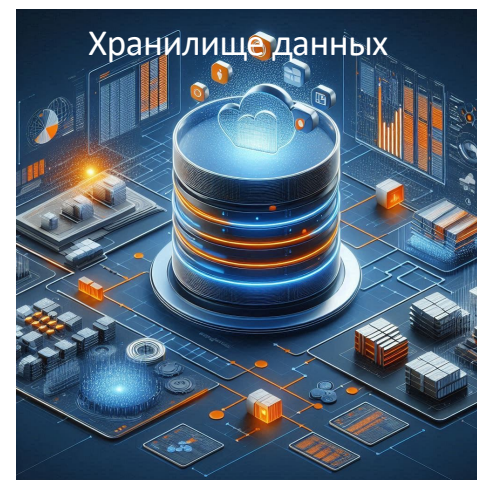


Технологические вызовы

- Множество разнообразных источников данных
- Разные протоколы, разные форматы
- Низкое качество данных (по стандартам статистики)
- Проблемы классификации и сопоставления

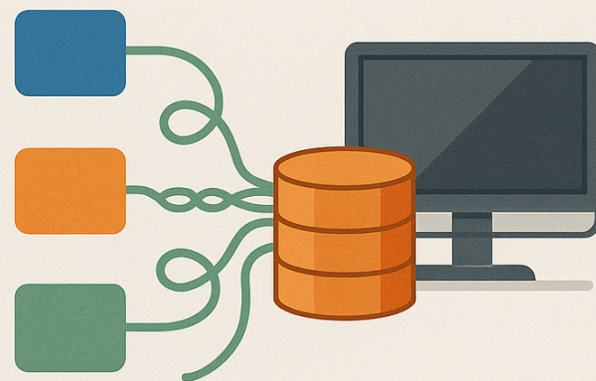
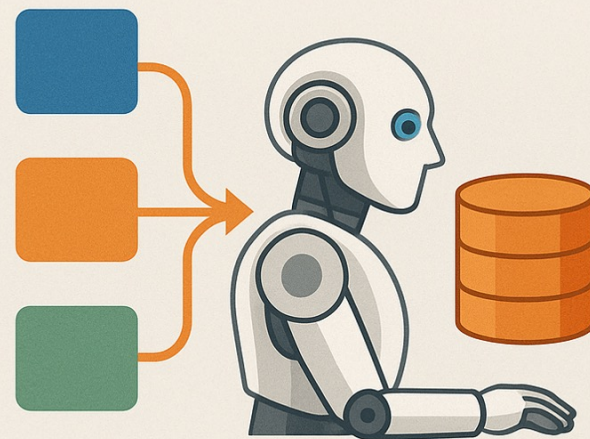


Методы сбора административных данных

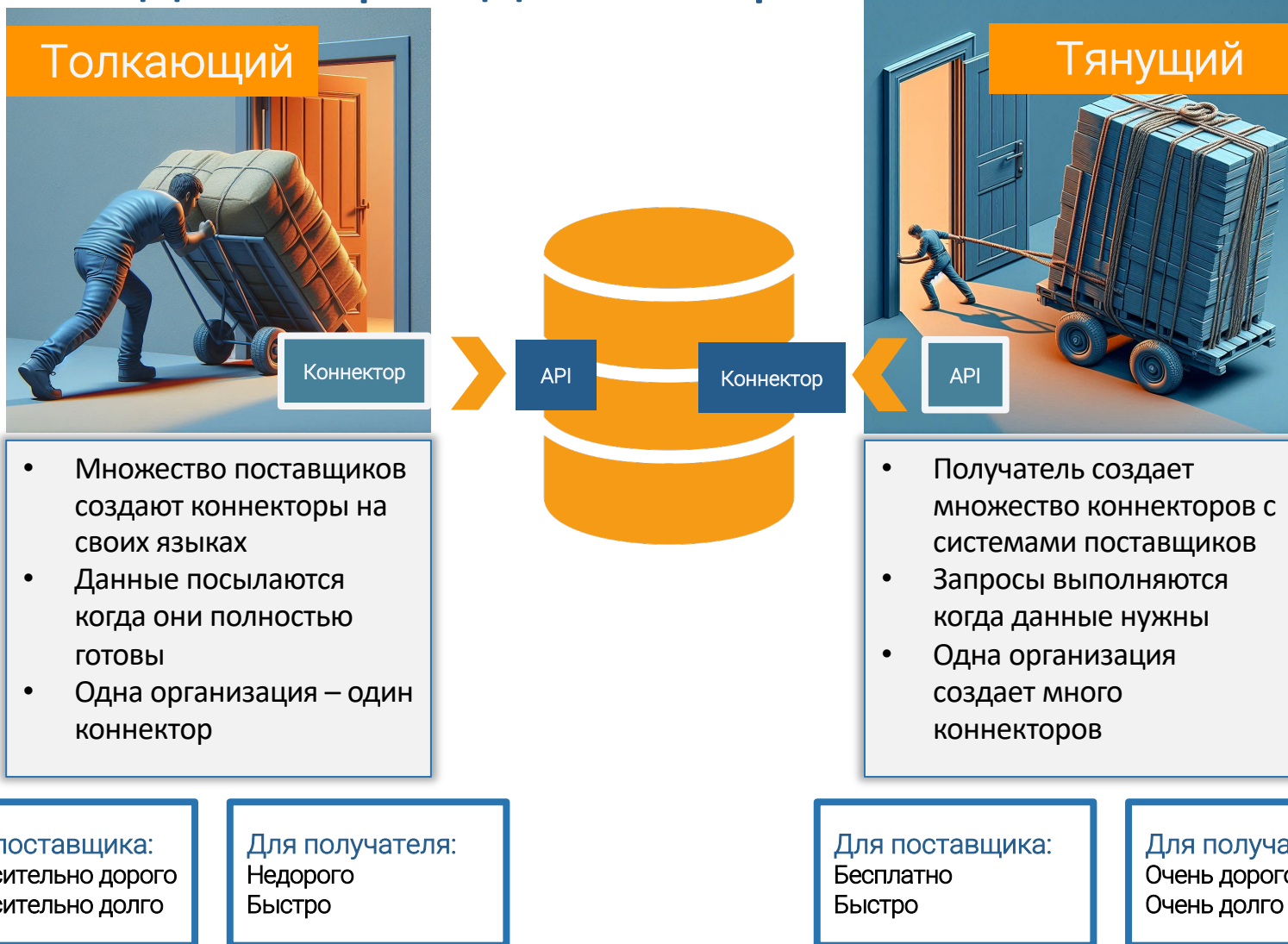


Способы предоставления административных данных

- Толкающий режим – поставщики сами отправляют данные в нашем формате в наш API
- Тянувший режим – мы ходим в API поставщиков и забираем данные в их формате



Методы сбора административных данных

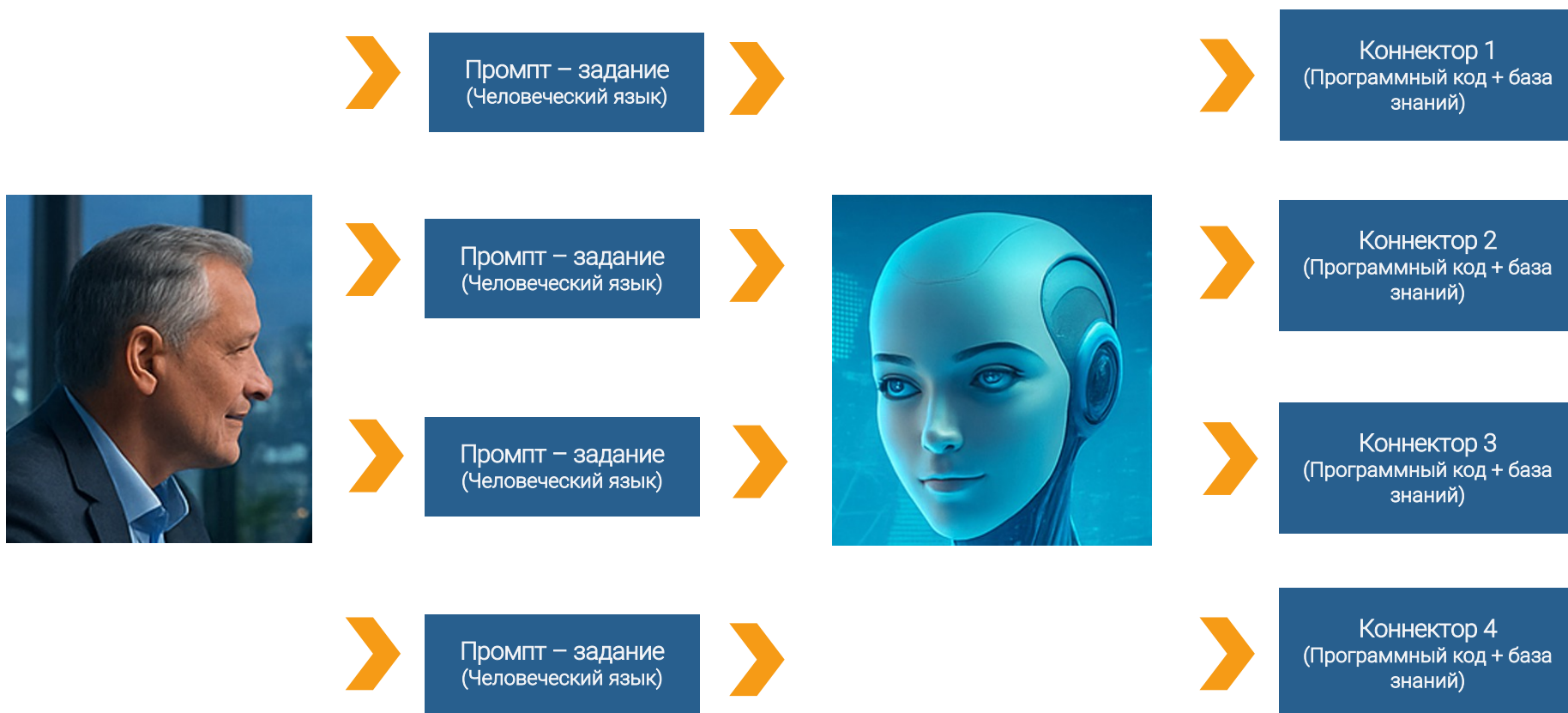


ИИ:Сбор и предварительная обработка данных

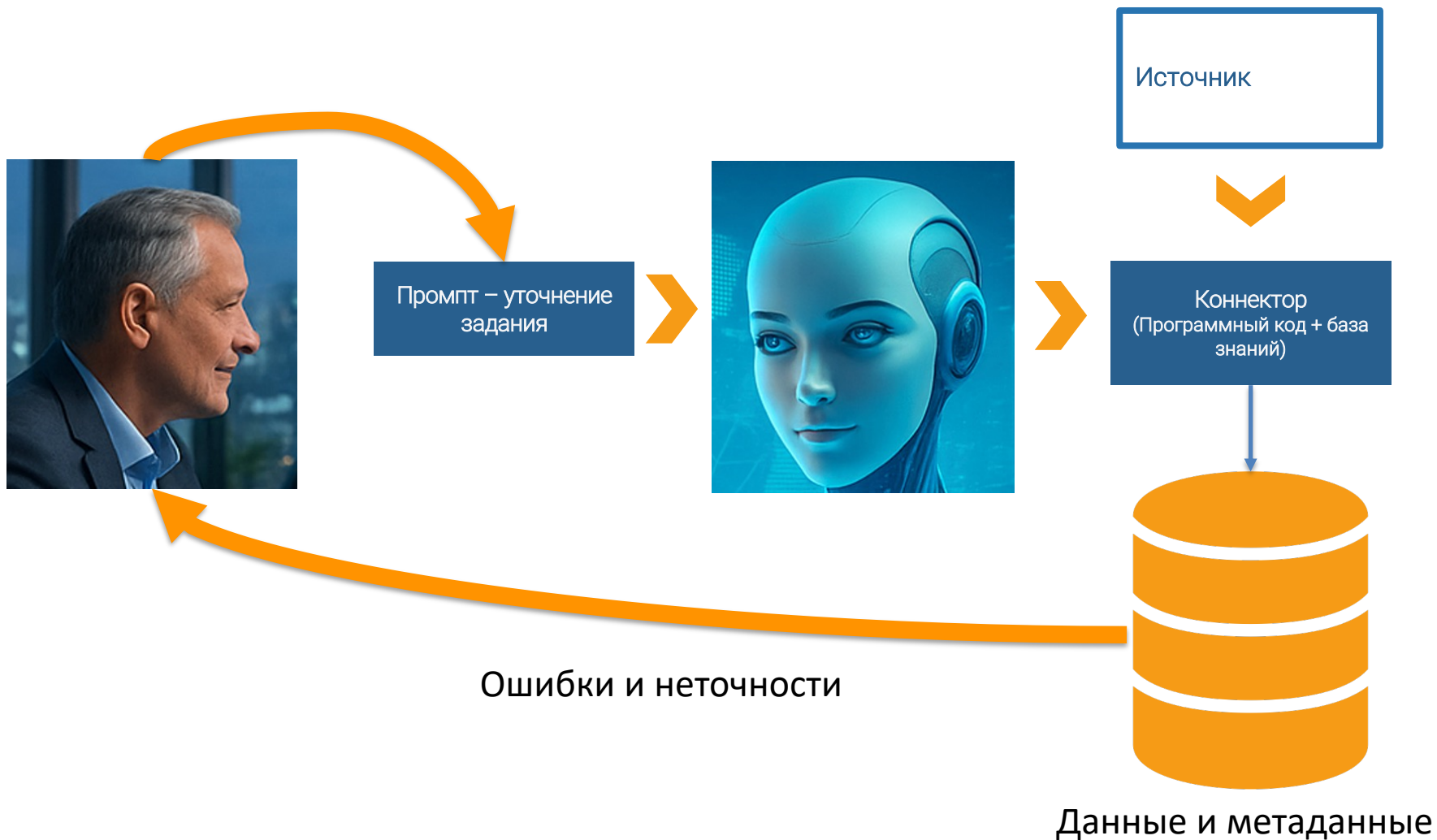
- Сверхбыстрая генерация тянущих коннекторов для любых API, FTP, облачных хранилищ
- Сверхбыстрая генерация конверторов из любых форматов во входной формат получателя
- Выявление, анализ и классификация ошибок, генерация журналов ошибок с рекомендациями по исправлению
- Семантическая обучаемая переклассификация в НСИ получателя
- Загрузка в целевую систему с использованием ее API и библиотек



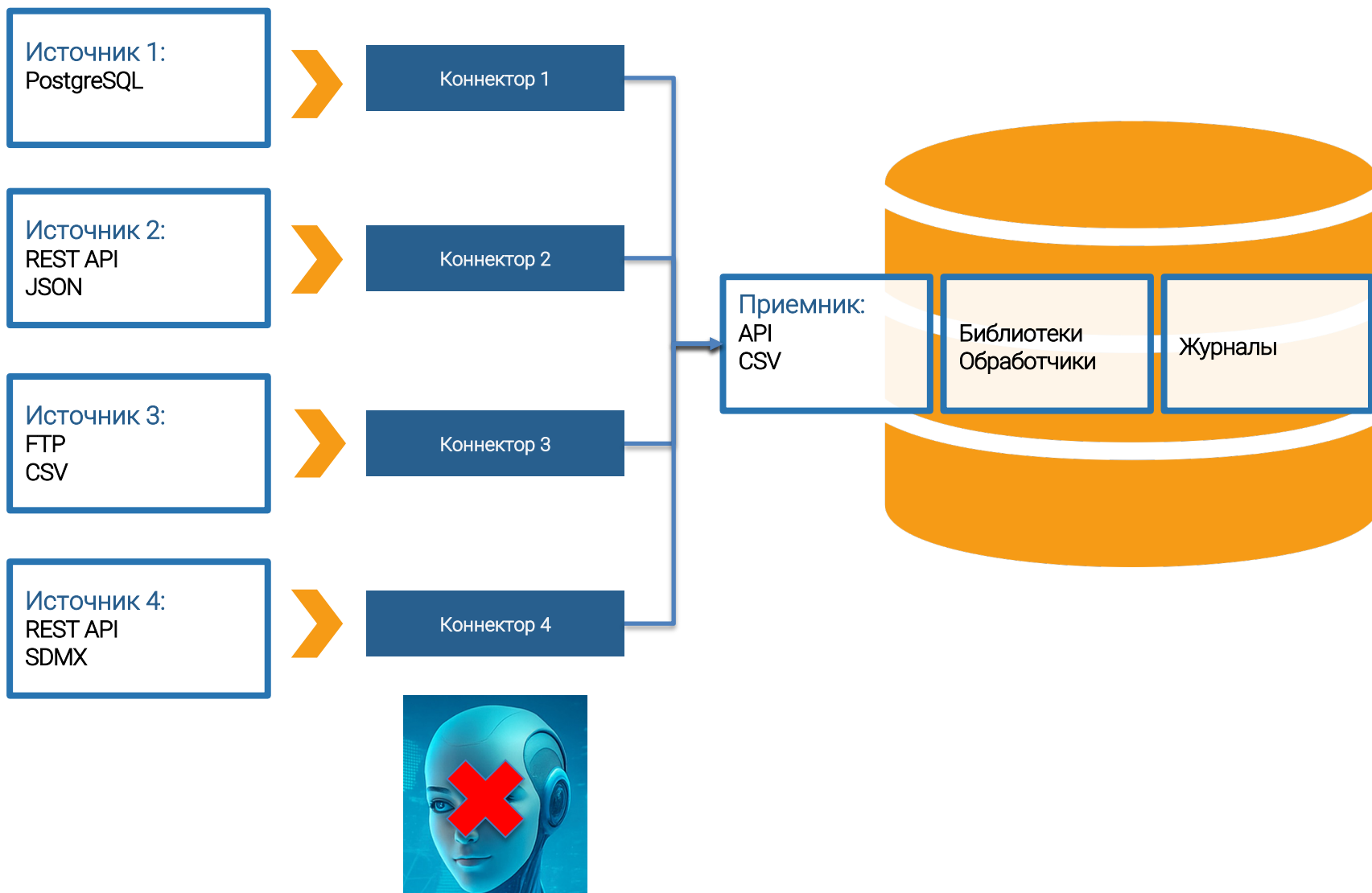
ИИ: Генерация отчуждаемых коннекторов



ИИ: Итерационная тренировка

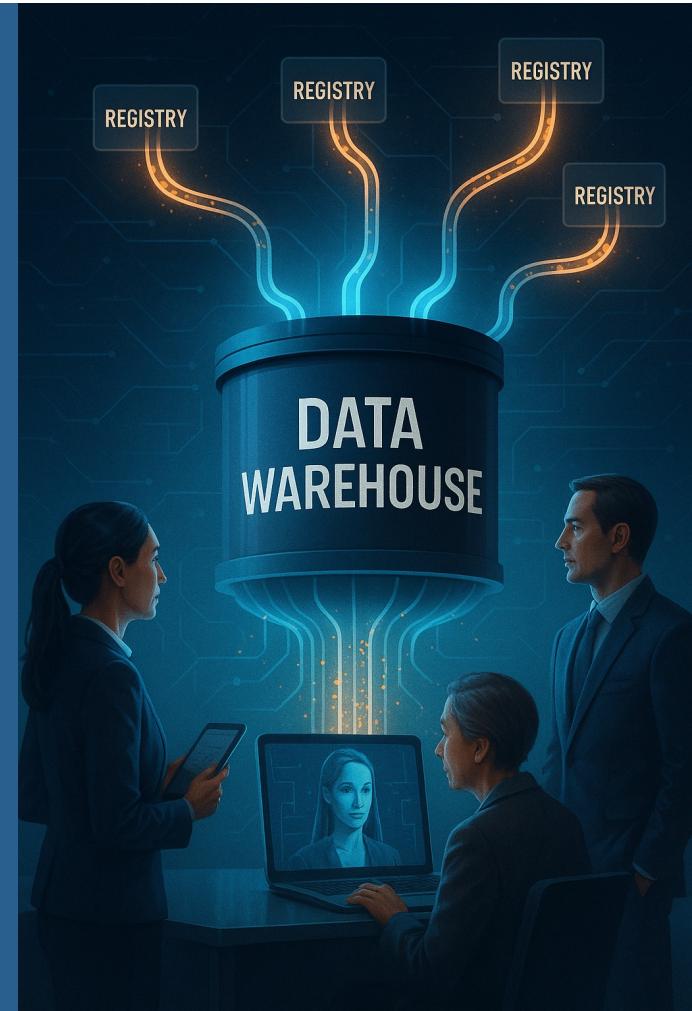


Сбор данных (с отключенным ИИ)



Валидация, исправление и обогащение данных

- Валидация и очистка данных:
 - Проверка корректности формата
 - Проверка полноты данных
 - Исправление формата (знакокодировки)
 - Семантическое распознавание и исправление имен полей
 - И так далее
- Гармонизация данных:
 - Замена текстов на коды
 - Замена кодов справочников на глобальные
- Обогащение данных:
 - Добавление атрибутов с вычислением их значений
 - Добавление вычисляемых полей
 - Связывание со статистическими массивами
- Генерация журналов ошибок с примерами ошибок и рекомендациями по исправлению



Расчет статистических показателей

- Генерация процедур расчета статистических показателей из полученных, очищенных и гармонизированных данных
- Сложные цепочки расчетов с условиями
- Использование OLAP для многомерных расчетов
- Проверка результатов по заданным алгоритмам
- Генерация журнала расчетов



Примеры проектов

Расчет индекса потребительских цен на основе данных авиакомпаний

- Генерация коннектора получающего десятки миллионов билетов от десятков авиакомпаний
- Валидация и исправление десятков типов ошибок
- Классификация ошибок, генерация журналов ошибок
- Расчеты среднегеометрических цен
- Импутация данных для заполнения пропусков
- Прочие расчеты

Примеры проектов

Сбор админданных и расчет показателей

- Валидация административных данных
- Распознавание текстовых описаний периодичности показателей, генерация формальных описаний
- Генерация календаря сбора данных и расчетов
- Исправление грамматических и синтаксических ошибок в справочниках и классификаторах

Спасибо за внимание