

# Using AI for processing of Administrative data



Vladimir Nekrasov, Contour Components LLC

# Prerequisites for the production of statistics from administrative data

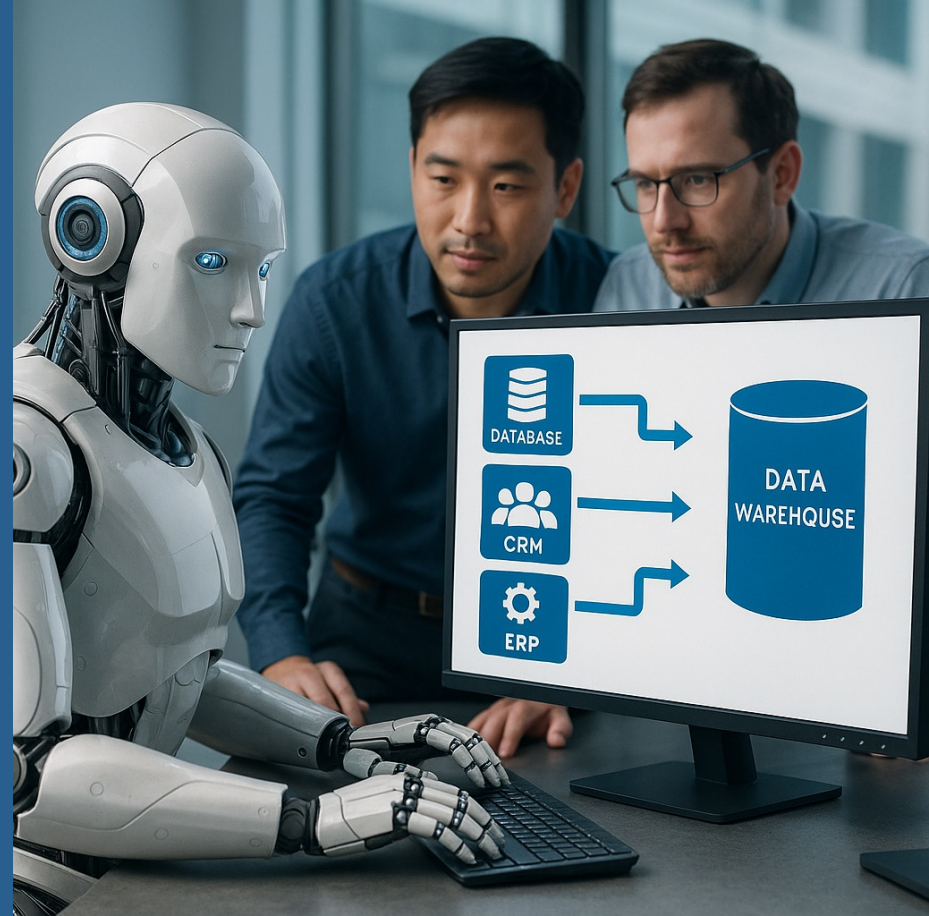
- Explosive growth of data in public administration
- Development of Information Technologies and Communications
- New Insights Features
  - Access to new sources
  - Improved accuracy
  - Increased agility



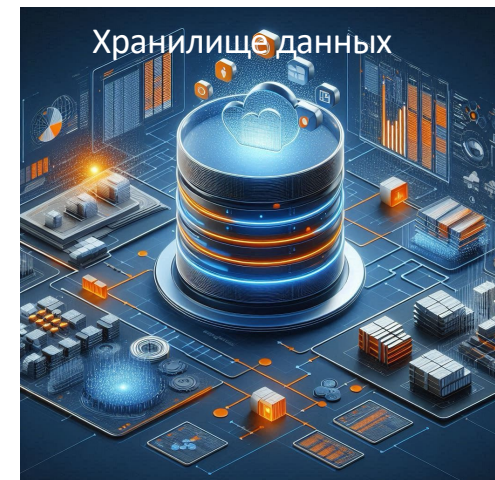
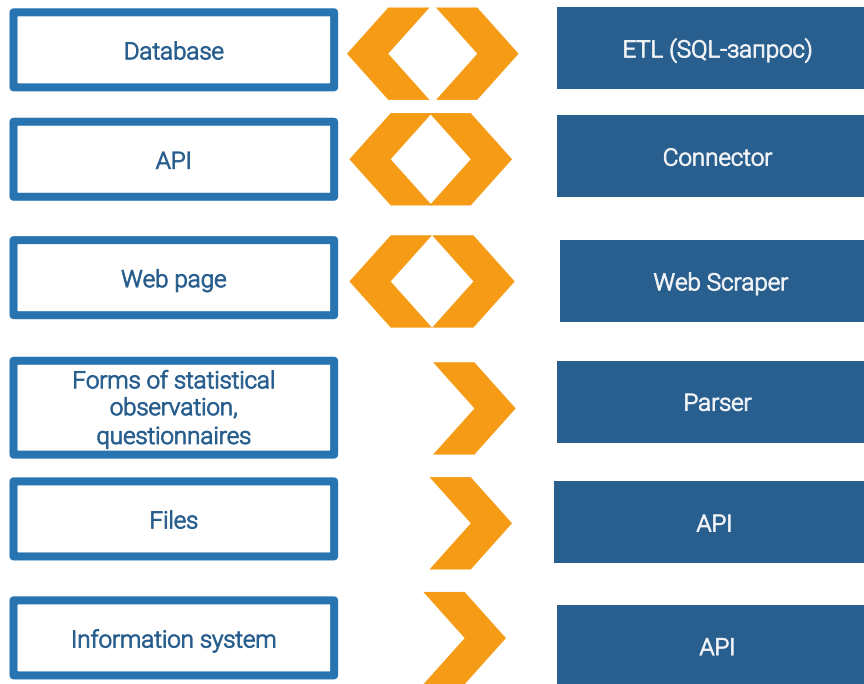


# Technological challenges

- A wide variety of data sources
- Different protocols, different formats
- Poor data quality (by statistical standards)
- Classification and Mapping Issues

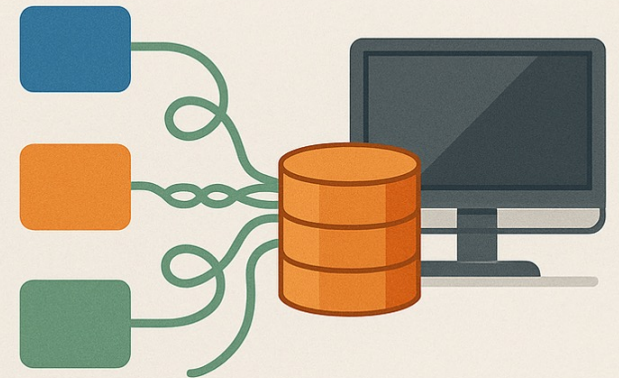
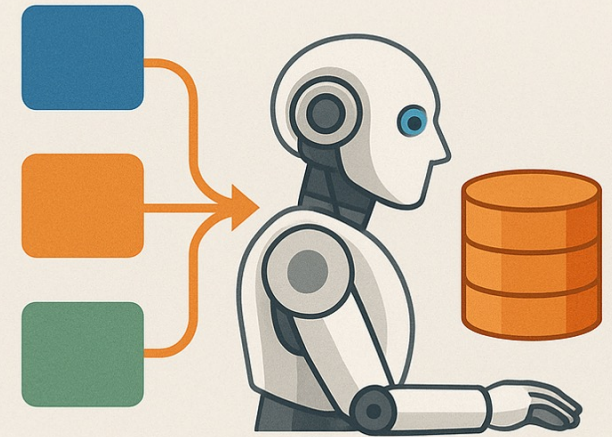


# Administrative data collection methods



# Methods of providing administrative data

- Push mode – suppliers themselves send data in our format to our API
- Pulling mode – we go to the API of suppliers and take data in their format



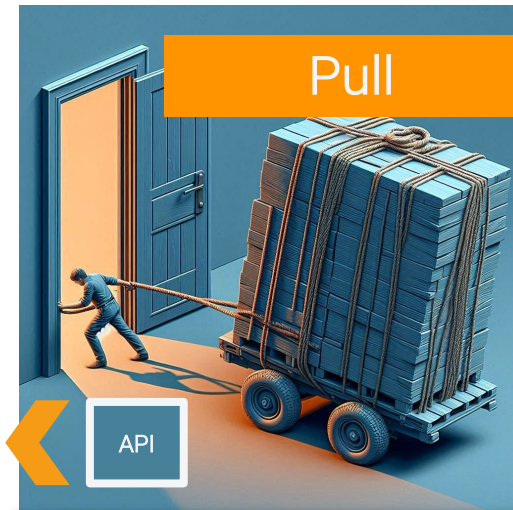
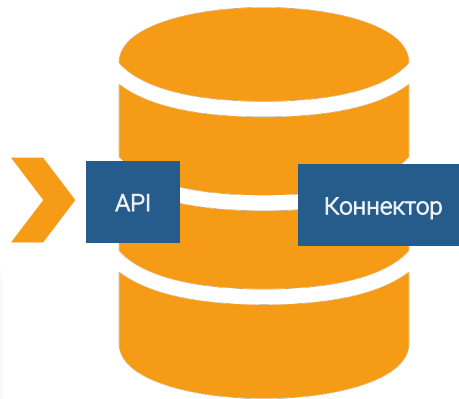
# Administrative data collection methods



- Many data providers create connectors in their own languages
- Data is sent when it is fully ready
- One organization – one connector

For the provider:  
Relatively expensive  
Relatively long

For the collector:  
Inexpensively  
Quickly



- The collector creates multiple connectors with providers systems
- Queries are executed when data is needed
- One organization creates many connectors

For the provider:  
Free  
Fast

For the collector:  
Really expensive  
Very long

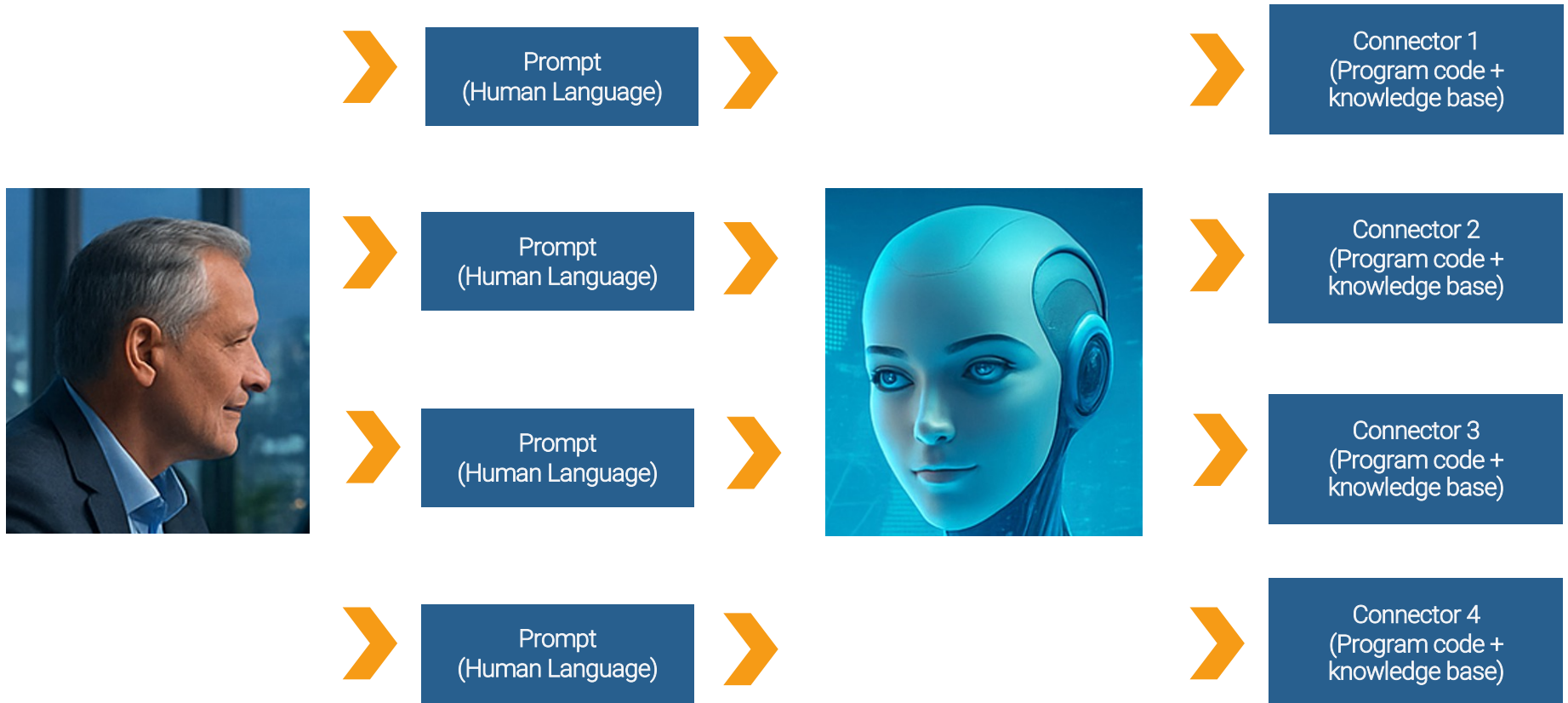


# AI:Data Collection & Pre-Processing

- Ultra-fast generation of pulling connectors for any API, FTP, cloud storage
- Ultra-fast generation of converters from any format to the input format of the recipient
- Identifying, analyzing and classifying errors, generating error logs with recommendations for correction
- Semantic Learnable Reclassification in the Recipient's Reference Data
- Loading into the target system using its APIs and libraries

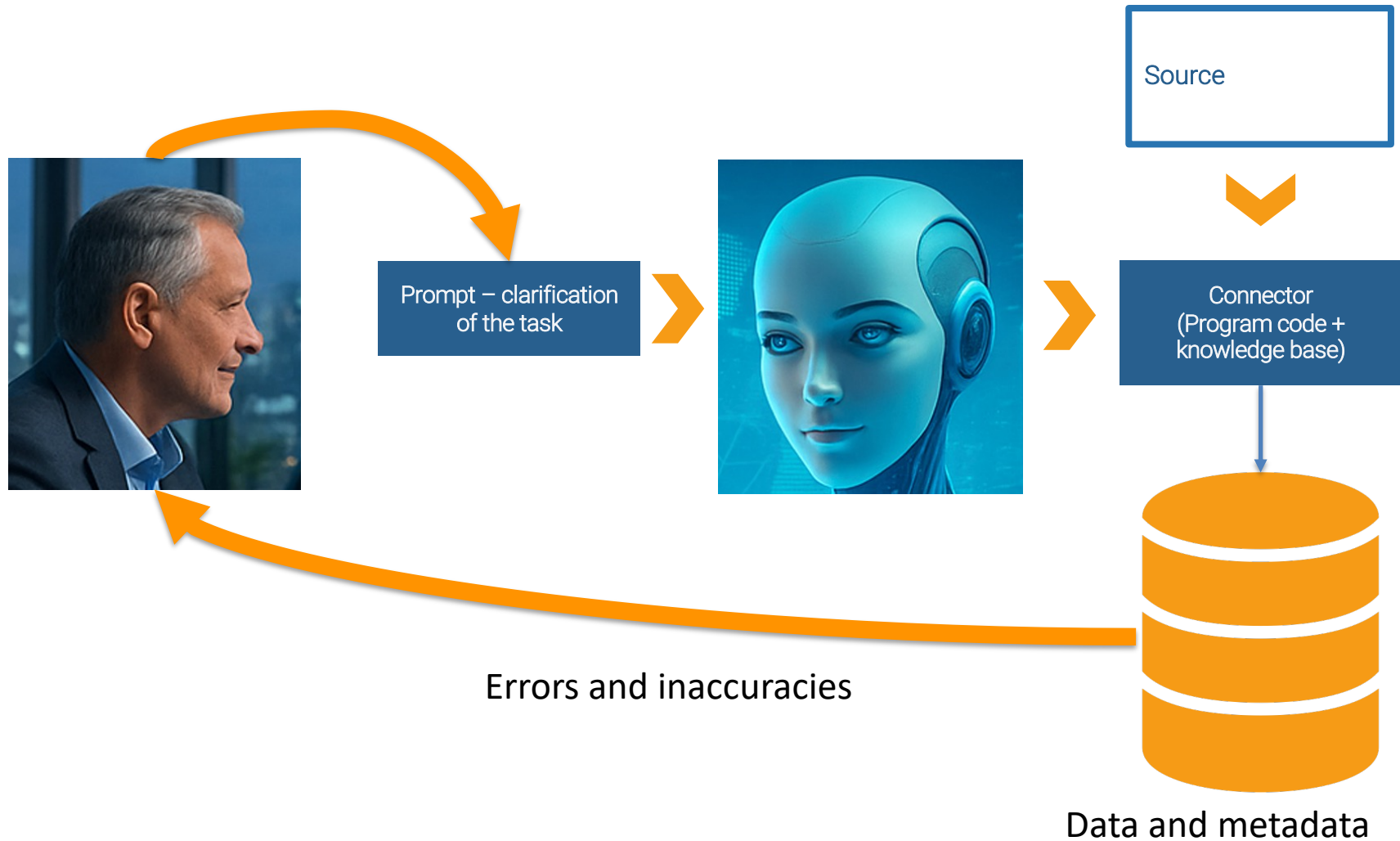


# AI: Generating Alienable Connectors

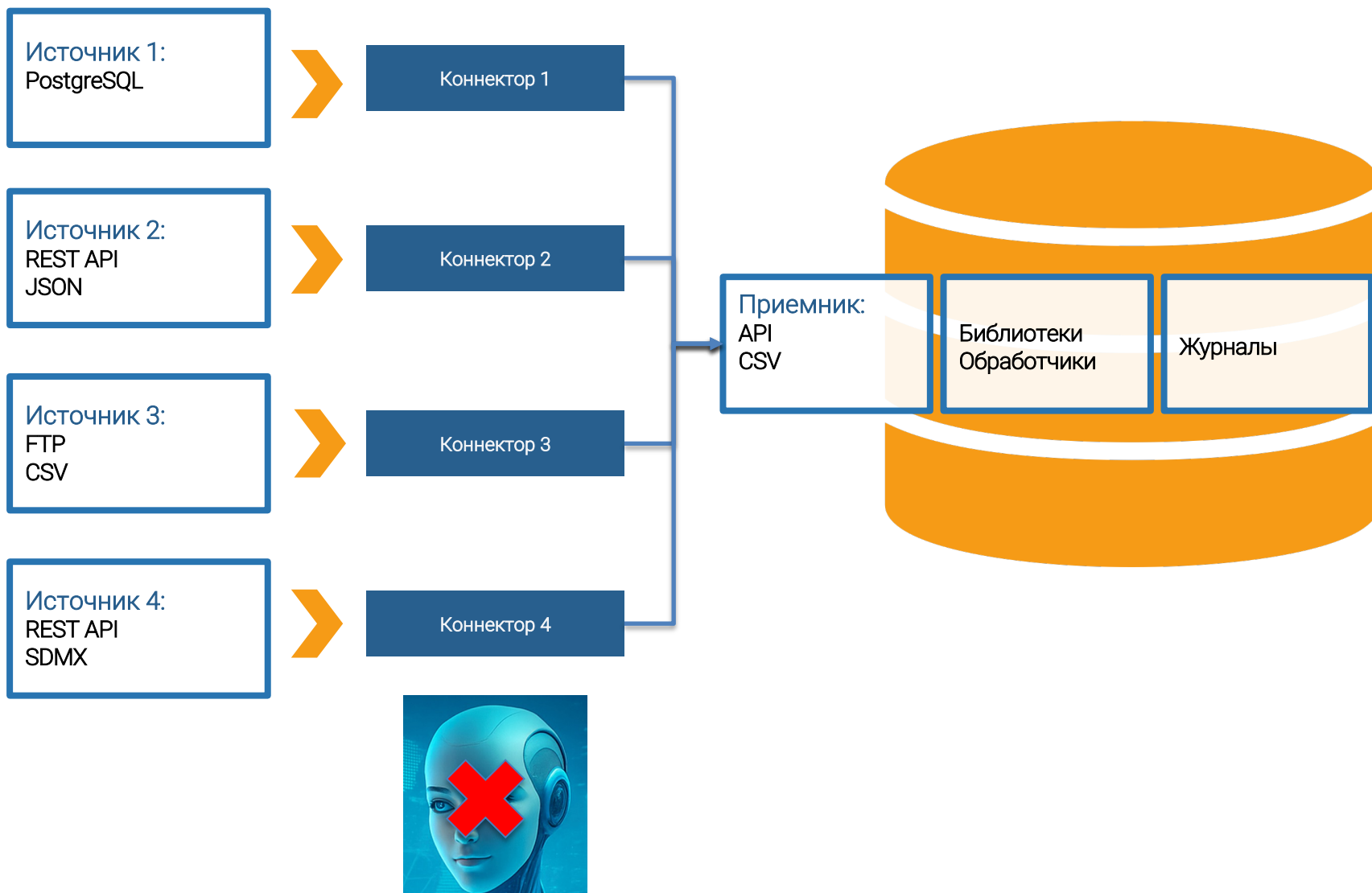




# AI: Iterative Training

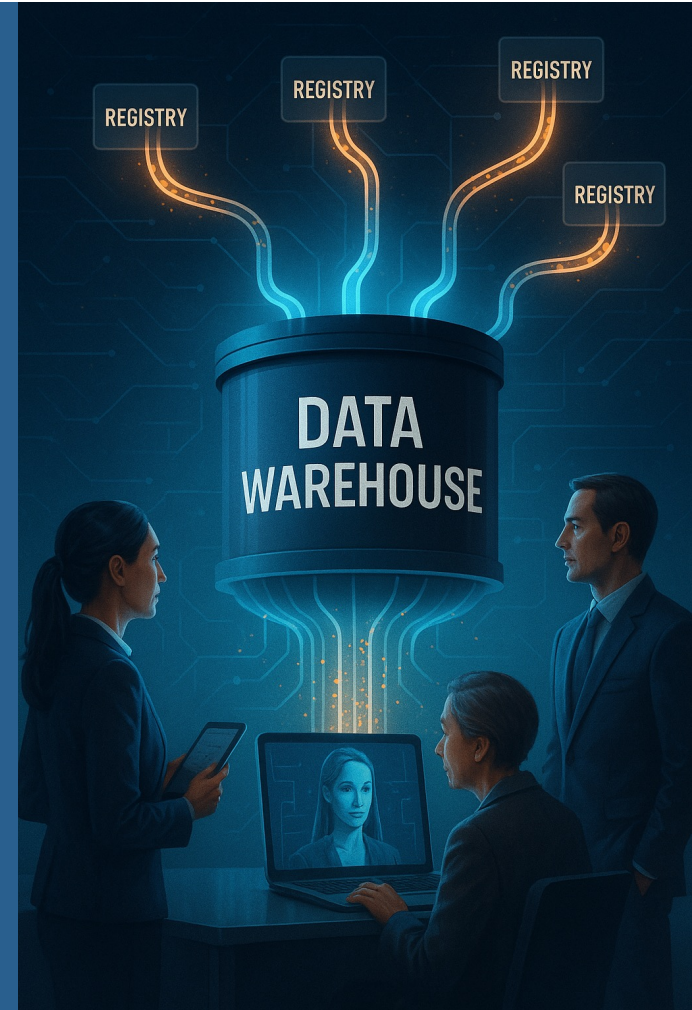


# Сбор данных (с отключенным ИИ)



# 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



# Calculation of statistical indicators and datasets

- Generation of procedures for calculating statistical indicators from obtained, cleaned and harmonized data
- Complex calculation chains with conditions
- Using OLAP for Multidimensional Calculations
- Checking the results according to the specified algorithms
- Generation of the calculation log





# Examples of projects

## Calculation of the consumer price index based on airline data

- Generation of a connector that obtaining tens of millions of tickets from dozens of airlines
- Validation **and correction** of dozens of error types
- Error classification, error log generation
- Geometric mean price calculations
- Imputation of data to fill in blanks
- Other calculations

# Examples of projects

## Collection of administrative data and calculation of indicators

- Validation of administrative data
- Recognition of text descriptions of the periodicity of indicators, generation of formal descriptions
- Generation of a calendar for data collection and calculations
- Correction of grammatical and syntactic errors in reference books and classifiers

Thank you for your  
attention