

# Deductor Data Quality – очистка персональных данных



**BaseGroup Labs**  
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ



# Качество персональных данных

# Состояние проблемы



# Типичная ситуация

Поле	Значение	Ошибка
Имя	Сергей	Первая буква - латинская
Фамилия	Петрович	Неверное поле
Страна	Лимония	Нет такой страны
Город	Мсква	Опечатка
Телефон	0000000	Фиктивный номер
E-mail	<a href="mailto:1@siteforspam.com">1@siteforspam.com</a>	Нет такого домена
Адрес	Новые Васюки 10	Нет адреса
Паспорт	АБВГДЕЖЗ	Не стандарт

## Издержки бизнеса

- Беспользый адресный маркетинг
- Потерянные контакты
- Отток клиентов
- Искажённая отчётность
- Неверные решения



# Решение: Deductor Data Quality

Deductor Data Quality – готовое решение для задачи повышения качества персональных данных:

- Оценка достоверности
- Очистка и систематизация
- Обогащение
- Дедупликация

## Схема работы





# ФИО: пример работы

## Разбить на составные части:

- Из ФИО выделить фамилию, имя, отчество

## Проверить каждый элемент

- Недопустимые символы
- Нет в справочниках
- Опечатки

## Оценить качество

- Просуммировать вес каждой ошибки
- Получить итоговый коэффициент доверия

## Исправить ошибки

- Исправить опечатки
- Привести к одному алфавиту
- Переставить спутанные поля

## Обогатить данные

- Заменить уменьшительные имена полными
- Определить пол

## Собрать из частей результат

- Собрать все элементы в ФИО по шаблону
- Сформировать результирующую запись

## Данные: Фамилия, имя, отчество

- Выделение фамилии, имени, отчества
- Проверка по справочникам с учетом ошибок
- Восстановление имени по сокращениям
- Определение пола

## Данные: Телефон

- Выделение нескольких телефонов из одного поля
- Выделение добавочных номеров
- Приведение к стандарту
- Определение городской/мобильный
- Определение оператора
- Определение страны, города

## Данные: E-mail

- Соответствие стандарту
- Выделение фамилии, имени, телефона, года рождения
- Регистрационные данные домена
- IP-адрес, существование и работоспособность домена
- Размещение сайта: страна, город, географические координаты

# Данные: Удостоверения личности

- Типы документов:
  - Паспорт
  - Водительское удостоверение,
  - Военный билет/удостоверение личности военнослужащего
  - Паспорт моряка
- Распознавание типа документа, страны, региона выдачи
- Страны: Россия, Украина, Белоруссия, СССР
- Исправление опечаток, проверка диапазонов
- Приведение к единому формату

## Данные: Адрес

- Разбиение на элементы
- Проверка по справочникам с учетом ошибок
- Проверка по КЛАДР, ФИАС
- Восстановление индекса, кода КЛАДР, ОКАТО

# Дедупликация

- Настройка стратегий дедупликации на базе комбинации элементов
- Дедупликация входного набора и всей базы персональных данных
- Формирование «золотой записи»
- Обогащение данных о клиенте из разных источников

# Следствия применения

## Мастер данные

- Формирование наиболее достоверных персональных данных
- Использование «золотой записи» о клиенте во всех системах компании

## Кросс-проверки

- Повышение доверия к информации
- Актуализация данных
- Обогащение данных



# Пример работы

№	Поле	Значение
1	<b>12</b> Индекс в группе	5
2	<b>ab</b> Вход - Код	a9228
3	<b>ab</b> Вход - Адрес	СВЕРДЛОВСКАЯ АСБЕСТ МИРА,6.1.57
4	<b>ab</b> Индекс	624260
5	<b>ab</b> Регион - Тип	ОБЛ
6	<b>ab</b> Регион - Название	СВЕРДЛОВСКАЯ
7	<b>ab</b> Район - Тип	
8	<b>ab</b> Район - Название	
9	<b>ab</b> Город - Тип	Г
10	<b>ab</b> Город - Название	АСБЕСТ
11	<b>ab</b> Поселение - Тип	
12	<b>ab</b> Поселение - Название	
13	<b>ab</b> Улица - Тип	УЛ
14	<b>ab</b> Улица - Название	МИРА
15	<b>ab</b> Код КЛАДР	66000002000002400
16	<b>ab</b> Дом	6
17	<b>ab</b> Корпус	1
18	<b>ab</b> Строение	
19	<b>ab</b> Кв. \Оф.	57
20	<b>ab</b> Помещение	
21	<b>ab</b> Нераспознанный остаток строки	
22	<b>12</b> Номер запроса	0

Было

Стало

Определен  
почтовый индекс

Определен тип  
населённого пункта

Код КЛАДР

Выделен номер  
квартиры



**BaseGroup Labs**  
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

# ВЫГОДЫ

## Режим использования

- Услуга: разовая или регулярная очистка
- Приложение: пакетная очистка
- On-Site: внутренний веб-сервис
- On-Demand: облачный веб-сервис

## Готовая логика

Не надо ничего настраивать, логика очистки встроена в веб-сервис



# Гарантированное качество



## Готовые справочники

<b>Сущность</b>	<b>Кол-во</b>
Доменных регионов	1 000
Имена (кириллица)	16 000
Фамилии (кириллица)	320 000
Имена (латиница)	20 000
Фамилии (латиница)	25 000
Операторы связи/регионы	160 000

## Срок внедрения

Первый результат за

**1** день



# Защита персональных данных



## Определения закона 152-ФЗ

**Персональные данные** - любая информация, относящаяся к прямо или косвенно определенному или определяемому физическому лицу (субъекту персональных данных).

**Обезличивание** персональных данных – действия, в результате которых становится невозможным без использования дополнительной информации определить принадлежность персональных данных конкретному субъекту персональных данных.

Статья 3, Федеральный закон РФ "О персональных данных" (152-ФЗ)

# Режим использования



Не требуется  
обезличивание

- Услуга
- Приложение
- On-Site

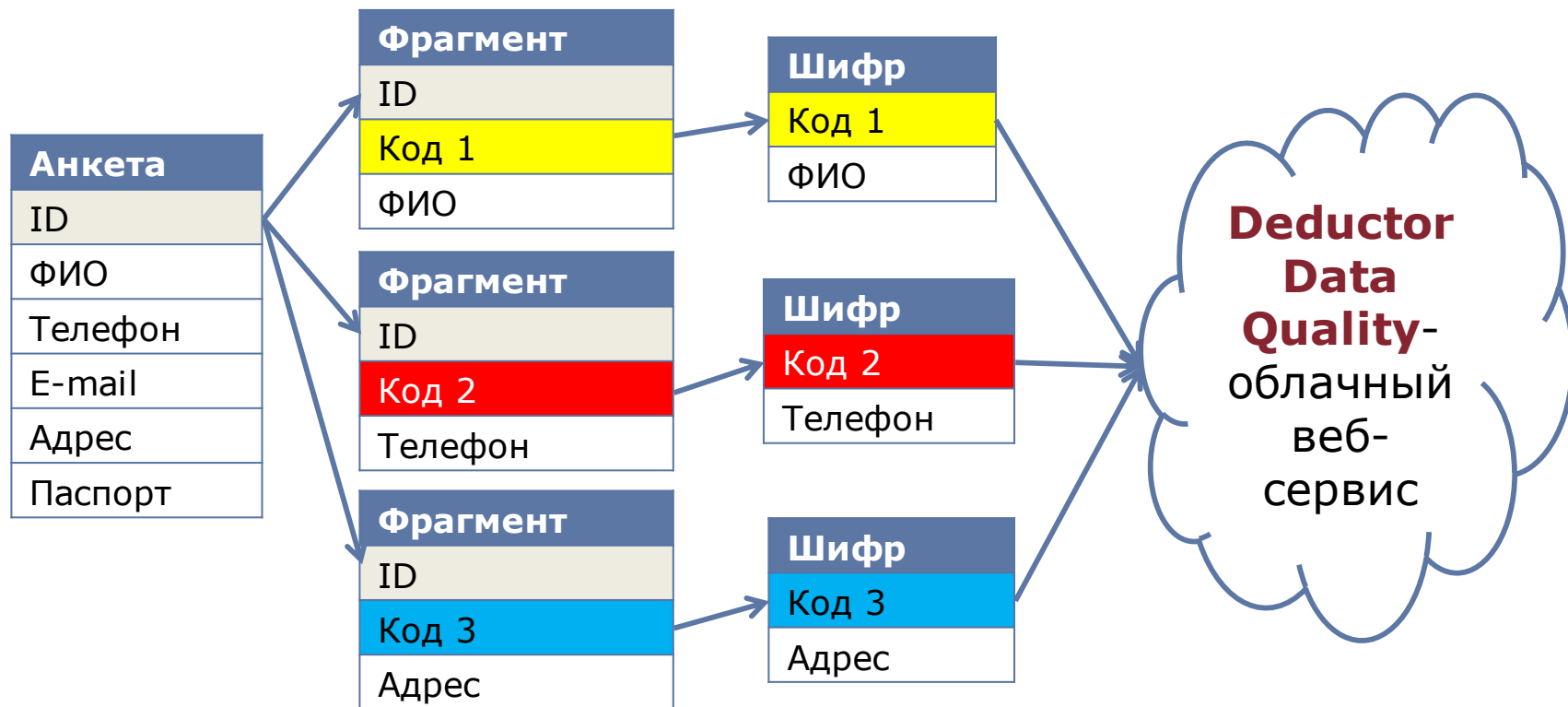


Требуется  
обезличивание

- On-Demand

Данные покидают  
пределы компании

# Обезличивание данных



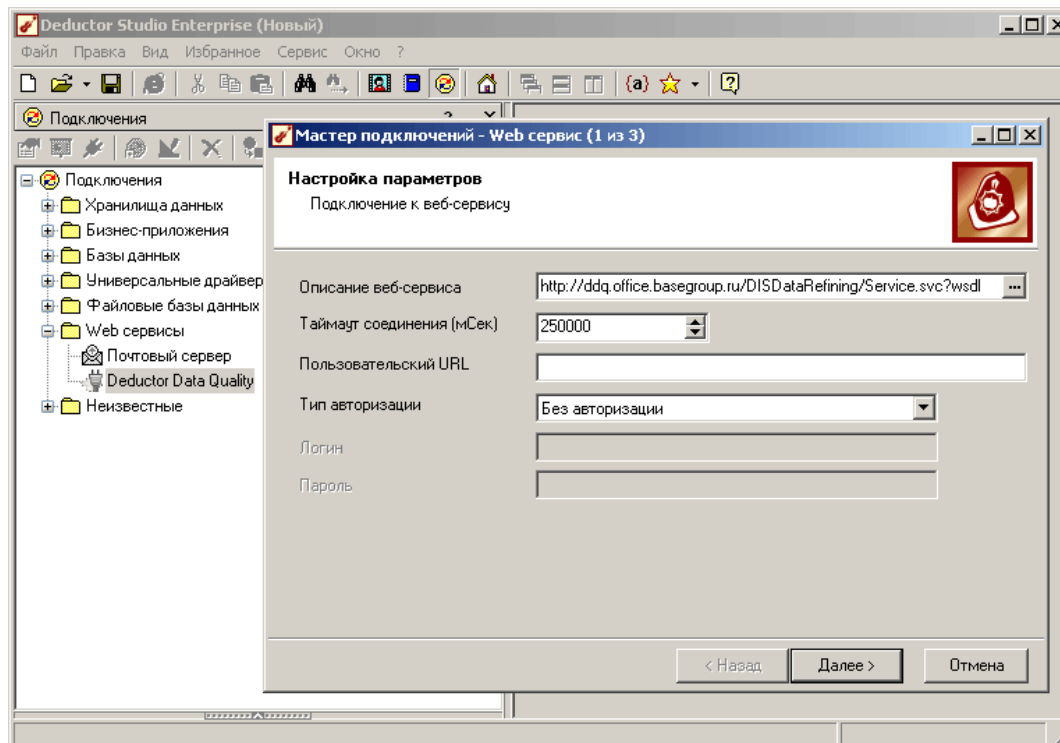
## Проблема дедубликации

После деперсонификации невозможно реализовать дедубликацию, т.к. необходим доступ всему объему данных.

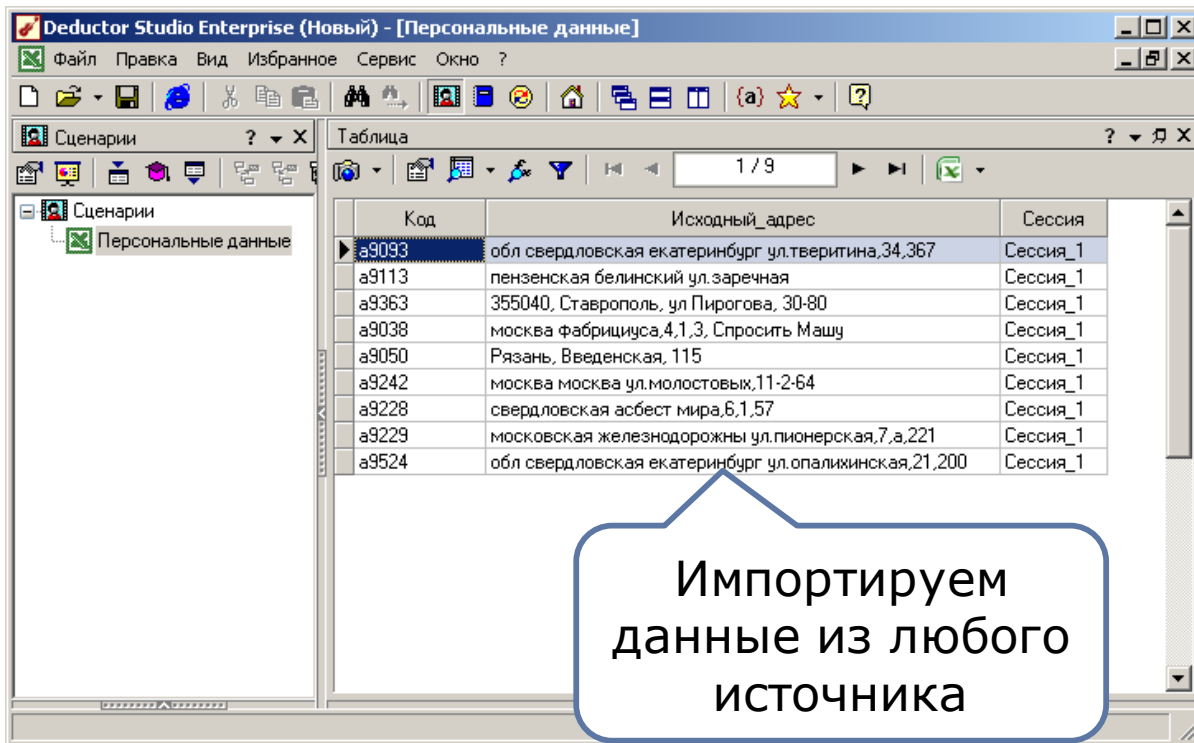
Дедубликацию рекомендуется производить внутри компании.

# Пошаговая демонстрация

# Шаг 1: Настроить подключение



## Шаг 2: Получить данные

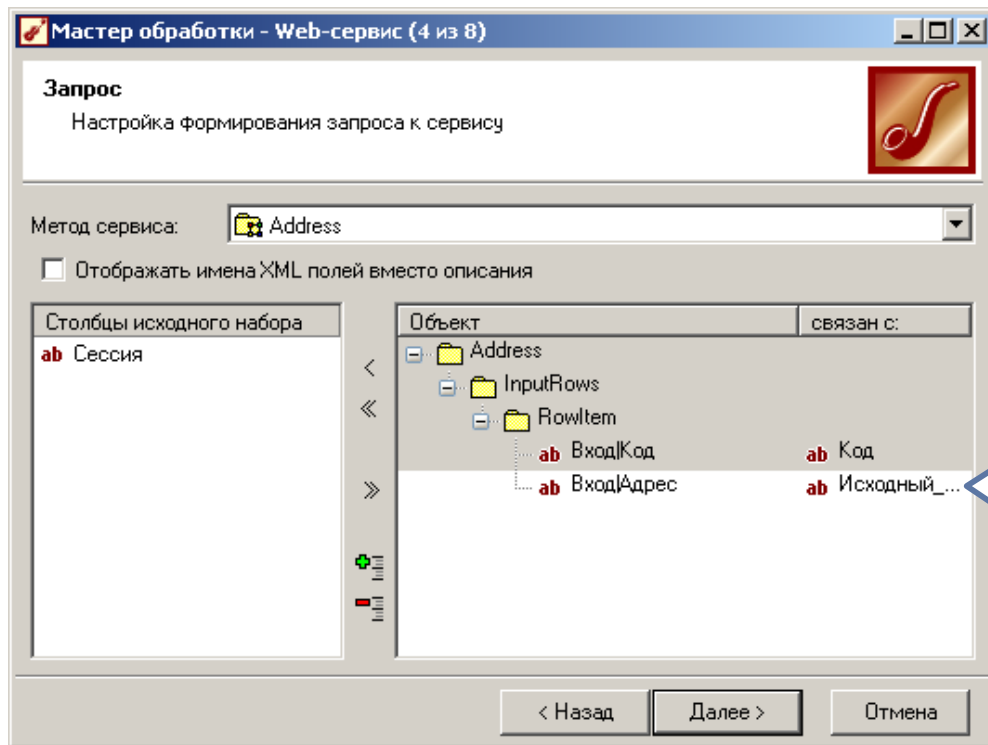


The screenshot shows the Deductor Studio Enterprise interface. The main window displays a table with the following data:

Код	Исходный_адрес	Сессия
a9093	обл свердловская екатеринбург ул.тверитина,34,367	Сессия_1
a9113	пензенская белинский ул.заречная	Сессия_1
a9363	355040, Ставрополь, ул Пирогова, 30-80	Сессия_1
a9038	москва фабрициуса,4,1,3, Спросить Машу	Сессия_1
a9050	Рязань, Введенская, 115	Сессия_1
a9242	москва москва ул.молостовых,11-2-64	Сессия_1
a9228	свердловская асбест мира,6,1,57	Сессия_1
a9229	московская железнодорожны ул.пионерская,7,а,221	Сессия_1
a9524	обл свердловская екатеринбург ул.опалихинская,21,200	Сессия_1

A callout box with a blue border and white background is positioned over the bottom right of the table, containing the text: "Импортируем данные из любого источника".

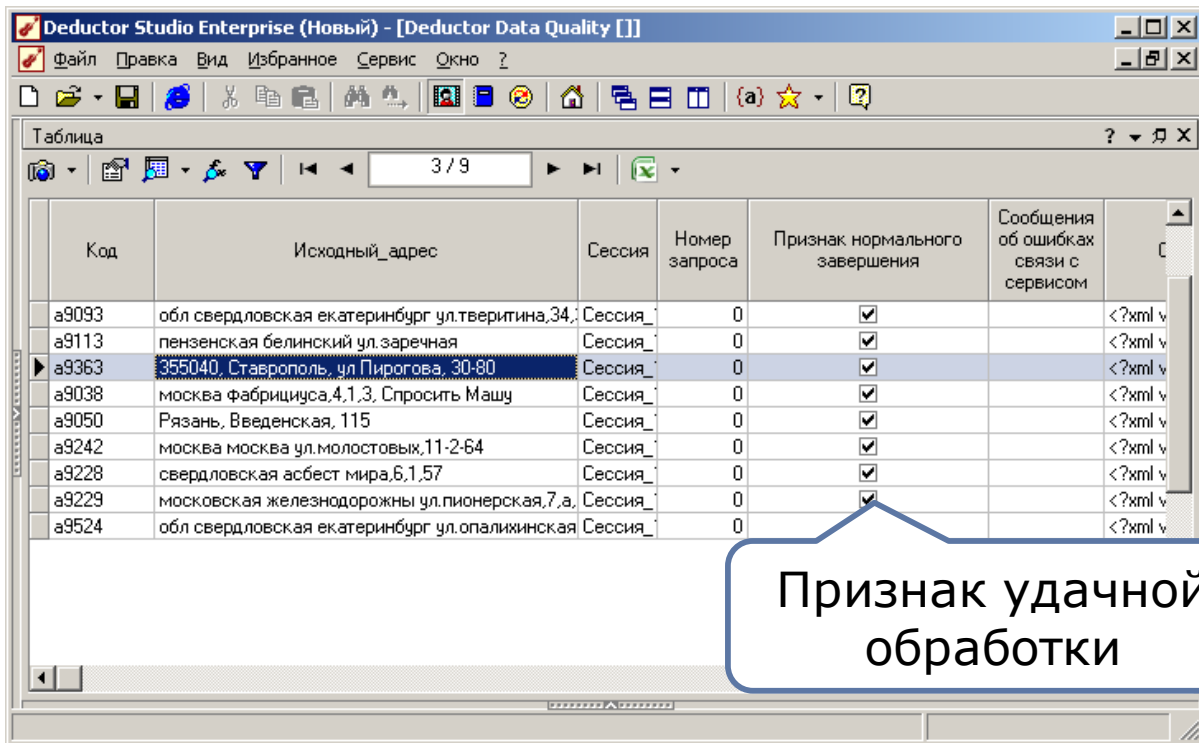
## Шаг 3: Настроить запрос



Настраиваем  
соответствие  
полей таблицы и  
сервиса



## Шаг 4: Получить ответ

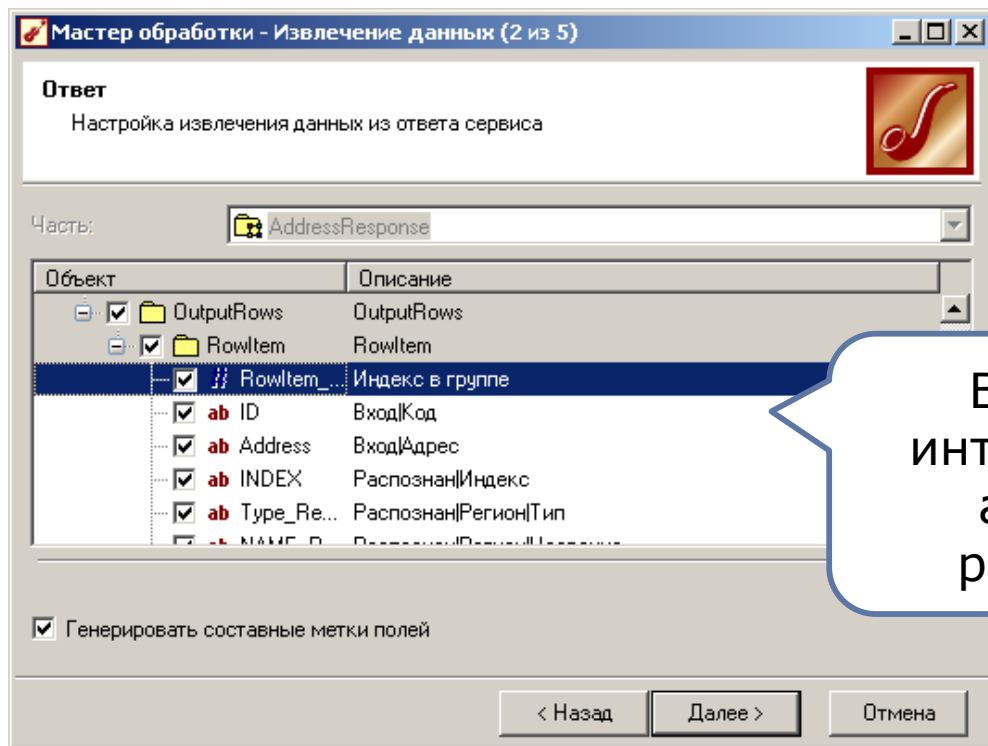


The screenshot shows the Deductor Studio Enterprise interface. The main window displays a table with the following columns: Код, Исходный\_адрес, Сессия, Номер запроса, Признак нормального завершения, and Сообщения об ошибках связи с сервисом. The table contains several rows of data, with the row for '355040, Ставрополь, ул. Пирогова, 30-80' highlighted. A callout bubble points to the 'Признак нормального завершения' column for this row, which contains a checked checkbox.

Код	Исходный_адрес	Сессия	Номер запроса	Признак нормального завершения	Сообщения об ошибках связи с сервисом
a9093	обл свердловская екатеринбург ул.тверитина,34,	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9113	пензенская белинский ул.заречная	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9363	355040, Ставрополь, ул. Пирогова, 30-80	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9038	москва фабрициуса,4,1,3, Спросить Машу	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9050	Рязань, Введенская, 115	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9242	москва москва ул.молостовых,11-2-64	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9228	свердловская асбест мира,6,1,57	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9229	московская железнодорожны ул.пионерская,7,а,	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v
a9524	обл свердловская екатеринбург ул.опалихинская	Сессия_	0	<input checked="" type="checkbox"/>	<?xml v

Признак удачной обработки

## Шаг 5: Разобрать ответ



Выбираем  
интересующие  
атрибуты  
результата

## Шаг 6: Получить результат

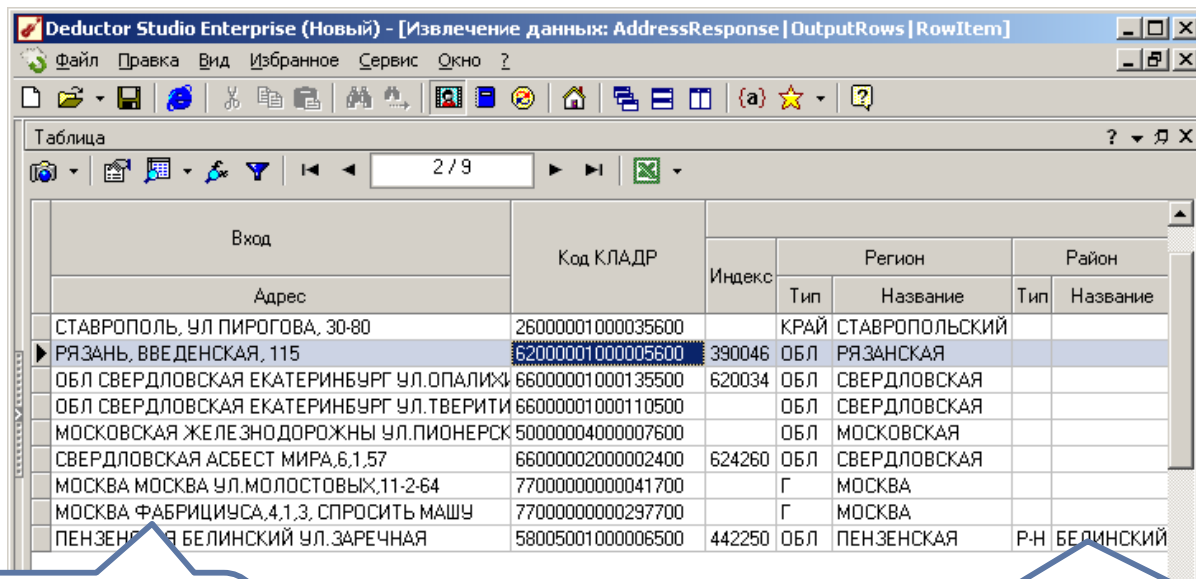


Table with 7 columns: Вход, Адрес, Код КЛАДР, Индекс, Регион (Тип, Название), Район (Тип, Название). The table contains 10 rows of data, with the second row highlighted in blue.

Вход	Адрес	Код КЛАДР	Индекс	Регион		Район	
				Тип	Название	Тип	Название
	СТАВРОПОЛЬ, УЛ ПИРОГОВА, 30-80	26000001000035600		КРАЙ	СТАВРОПОЛЬСКИЙ		
▶	РЯЗАНЬ, ВВЕДЕНСКАЯ, 115	62000001000005600	390046	ОБЛ	РЯЗАНСКАЯ		
	ОБЛ СВЕРДЛОВСКАЯ ЕКАТЕРИНБУРГ УЛ.ОПАЛИХИ	66000001000135500	620034	ОБЛ	СВЕРДЛОВСКАЯ		
	ОБЛ СВЕРДЛОВСКАЯ ЕКАТЕРИНБУРГ УЛ.ТВЕРИТИ	66000001000110500		ОБЛ	СВЕРДЛОВСКАЯ		
	МОСКОВСКАЯ ЖЕЛЕЗНОДОРОЖНЫ УЛ.ПИОНЕРСК	50000004000007600		ОБЛ	МОСКОВСКАЯ		
	СВЕРДЛОВСКАЯ АСБЕСТ МИРА,6,1,57	66000002000002400	624260	ОБЛ	СВЕРДЛОВСКАЯ		
	МОСКВА МОСКВА УЛ.МОЛОСТОВЫХ,11-2-64	77000000000041700		Г	МОСКВА		
	МОСКВА ФАБРИЦИУСА,4,1,3. СПРОСИТЬ МАШУ	770000000000297700		Г	МОСКВА		
	ПЕНЗЕНСКИЙ РАЙОН БЕЛИНСКИЙ УЛ. ЗАРЕЧНАЯ	58005001000006500	442250	ОБЛ	ПЕНЗЕНСКАЯ	Р-Н	БЕЛИНСКИЙ

Исходные  
«грязные»  
данные

Очищенные,  
стандартизированные и  
обогащенные данные



**BaseGroup Labs**  
ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

[basegroup.ru](http://basegroup.ru)