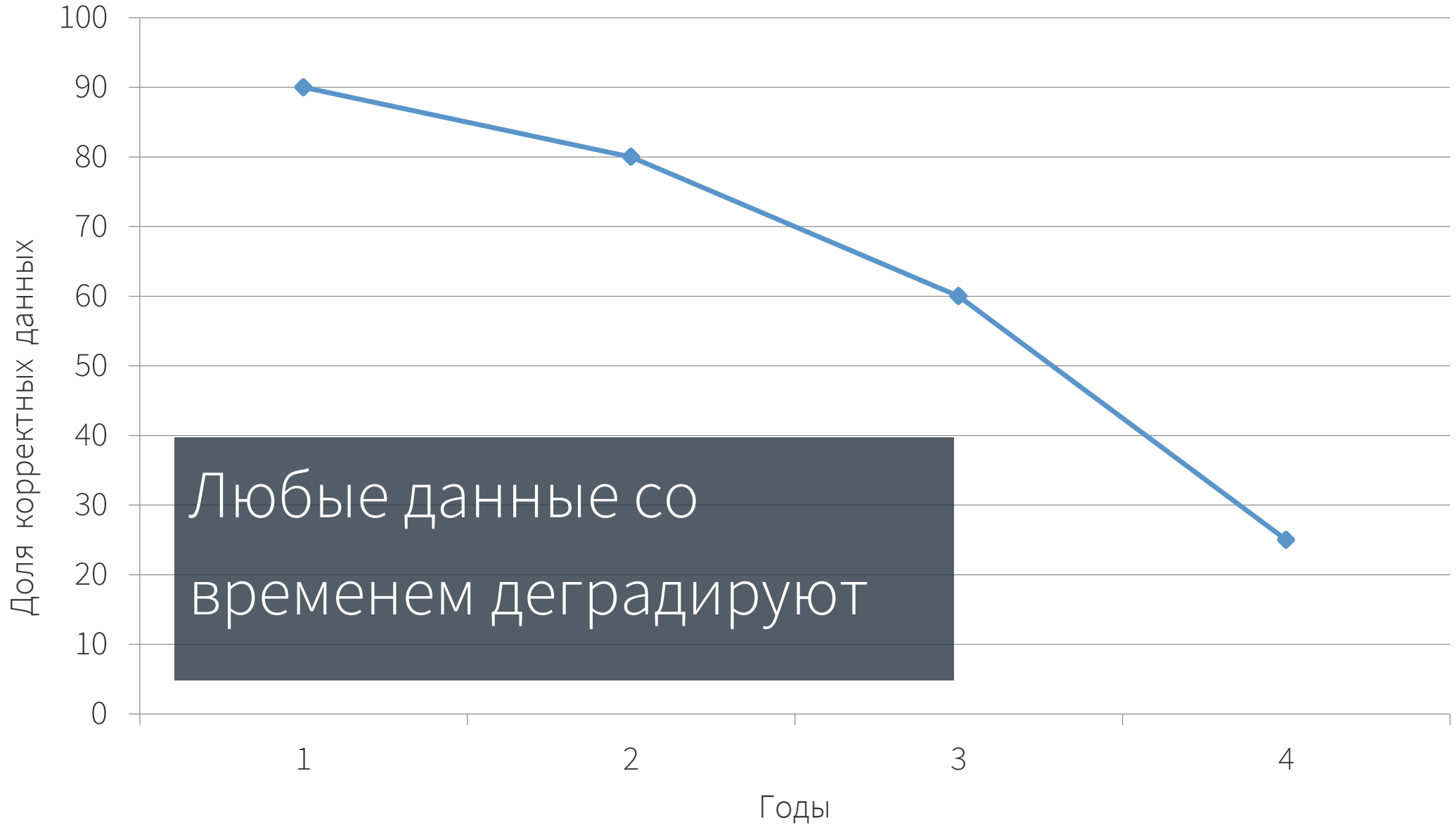
Loginom Data Quality: очистка данных

Люди ошибаются, поэтому проблемы с качеством данных будут всегда:

- Пропуски
- Опечатки
- Дубли и противоречия
- Фиктивные сведения
- Устаревшие данные


100
90
80
70
60
50
40
30
20
10
0
Доля корректных данных
1
2
3
4
Годы
Любые данные со временем деградируют

Последствия:

- Безадресный маркетинг
- Раздражающие коммуникации
- Потерянные контакты
- Отток клиентов
- Искаженная отчетность

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	1@siteforspam.com	Нет такого домена
Адрес	Новые Васюки 10	Нет адреса
Паспорт	АБВГДЕЖЗ	Не стандарт

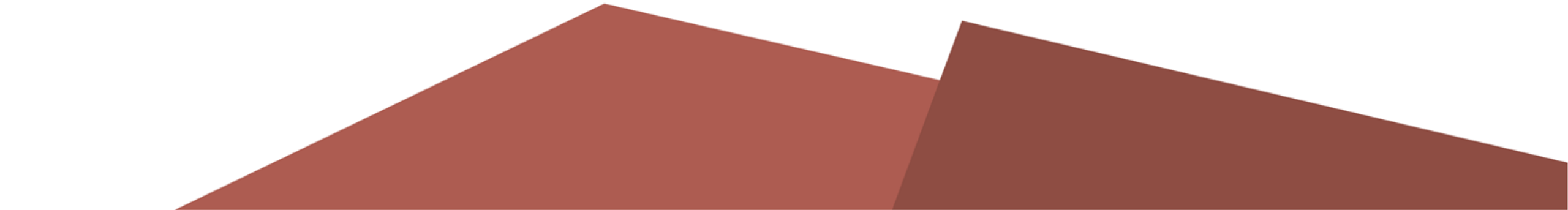
Решение Loginom Data Quality

Автоматически исправляет ошибки, приводит к стандартному виду, восстанавливает и обогащает клиентские данные

Задача	Описание
Очистка	Исправление ошибок, опечаток, фиктивных данных
Обогащение	Дополнение полезными данными, заполнение пропусков, актуализация
Стандартизация	Унификация представления данных
Дедупликация	Обнаружение дублей и «похожих» записей, объединение их в группы
Формирование «золотой записи»	Создание профиля клиента, включающей актуальные данных всех систем *

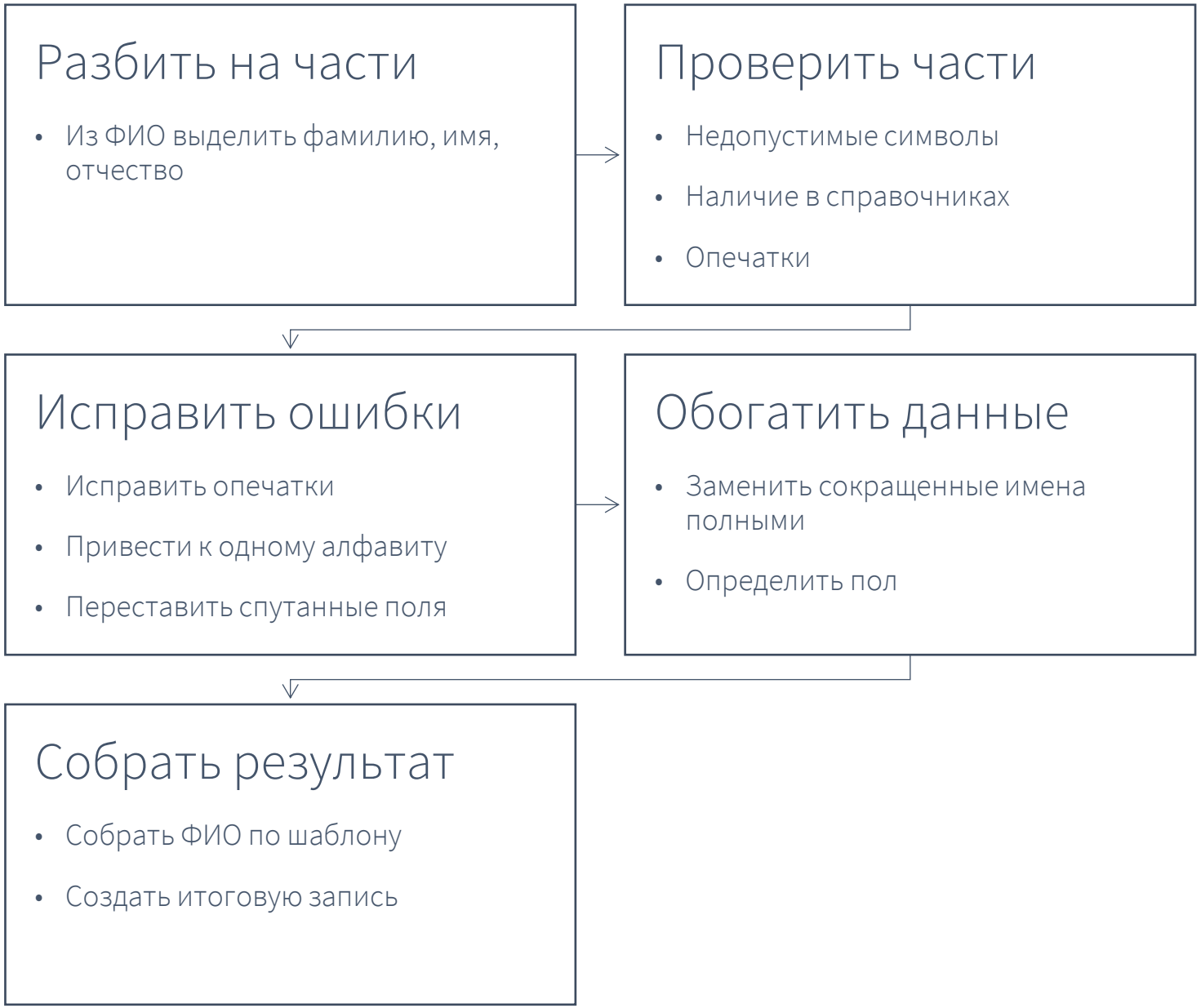
* В рамках проекта

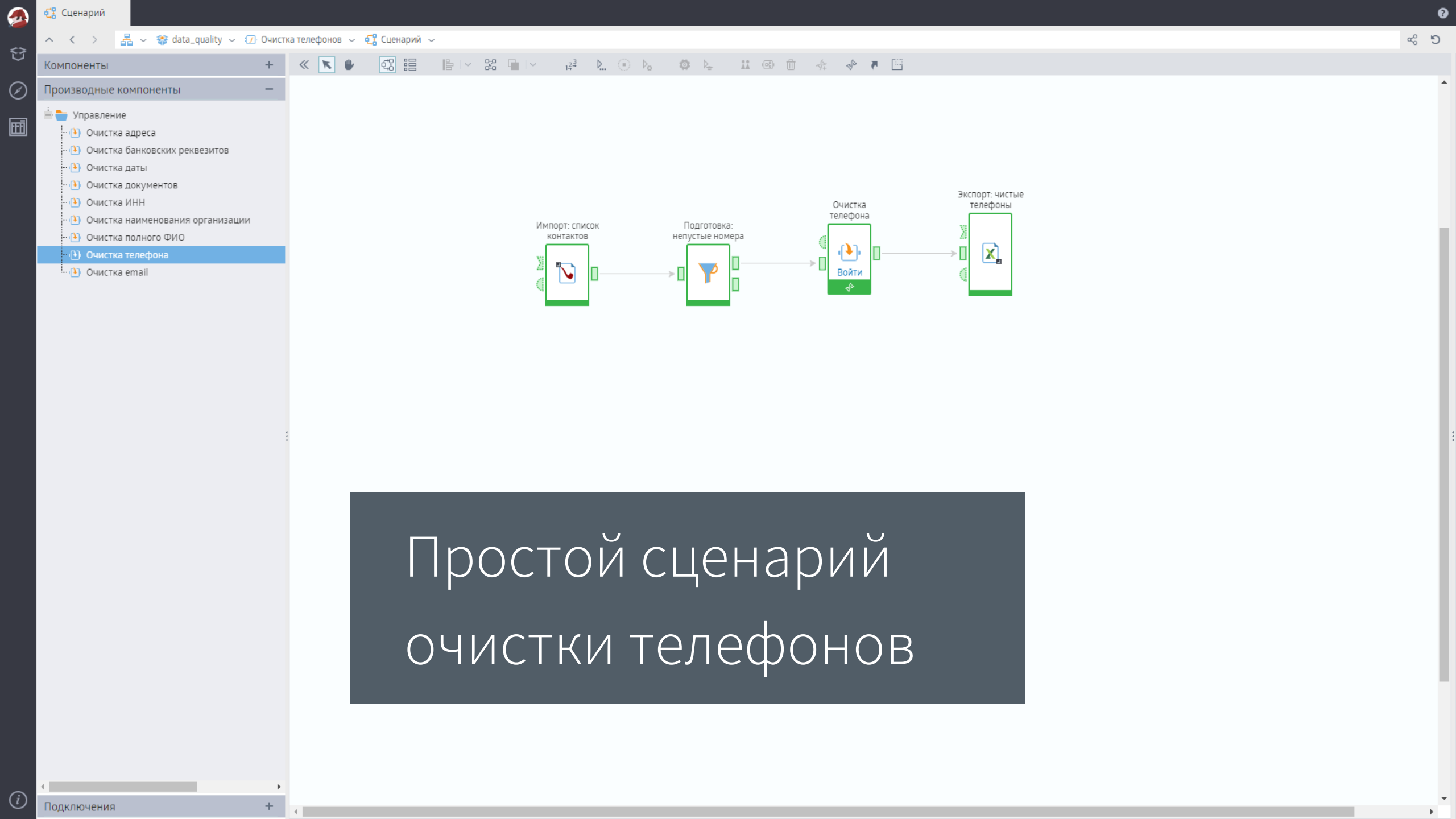
Очистка, обогащение и стандартизация

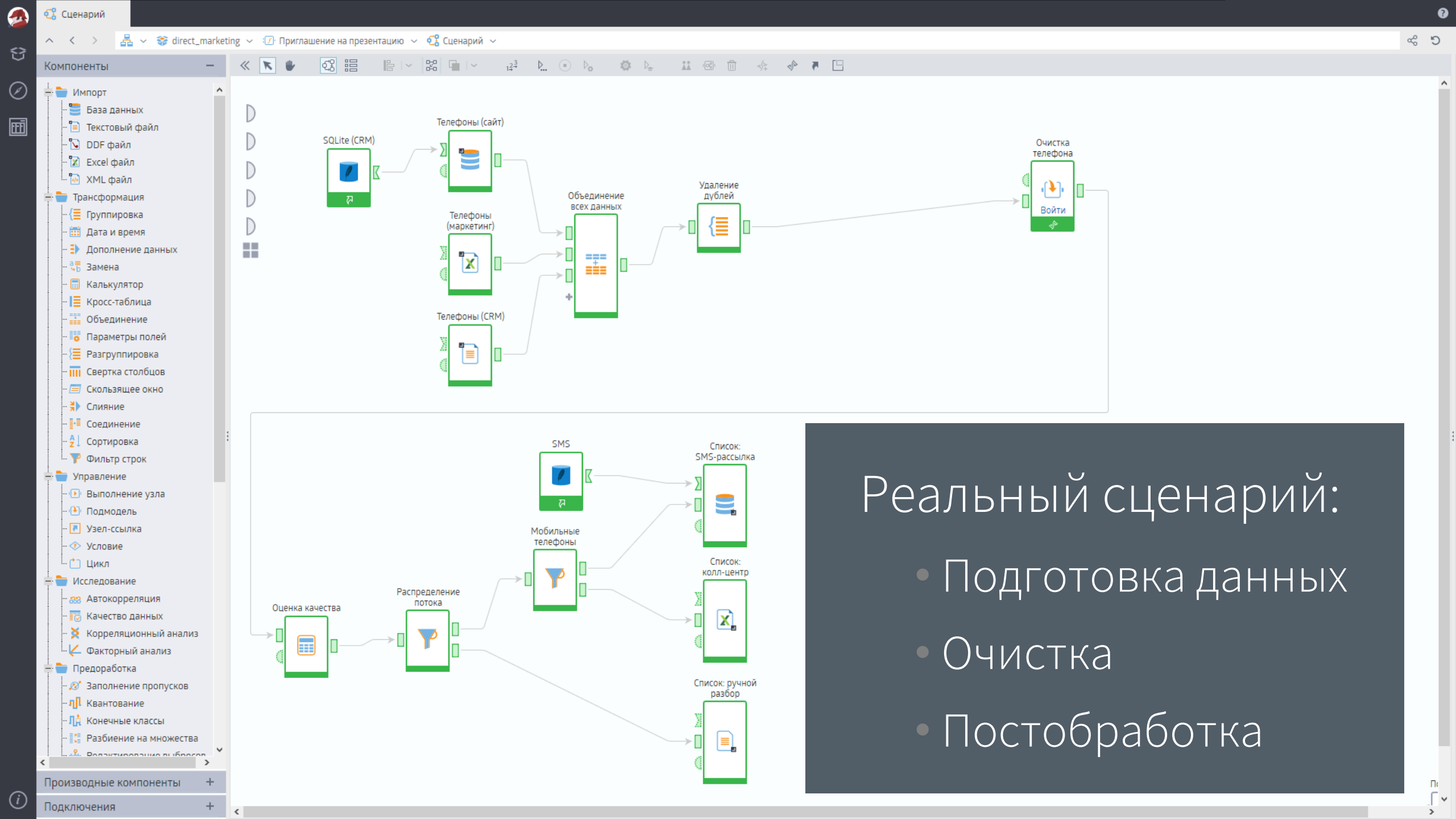


Очищаемые домены	Физ. лица	Юр. лица
Фамилия, имя, отчество	●	
Название организации		●
Почтовый адрес	●	●
Телефоны	●	●
Электронная почта	●	●
Удостоверения личности	●	
Реквизиты организаций		●
Даты	●	●
Банковские реквизиты		●

Пример схемы очистки ФИО




Сценарий
data_quality
Очистка телефонов
Сценарий
Компоненты
Производные компоненты
Управление
Очистка адреса
Очистка банковских реквезитов
Очистка даты
Очистка документов
Очистка ИНН
Очистка наименования организации
Очистка полного ФИО
Очистка телефона
Очистка email
Подключения
Импорт: список контактов
Подготовка: непустые номера
Очистка телефона
Войти
Экспорт: чистые телефоны
Простой сценарий очистки телефонов


Сценарий
direct_marketing
Приглашение на презентацию
Сценарий
Компоненты
Импорт
База данных
Текстовый файл
DDF файл
Excel файл
XML файл
Трансформация
Группировка
Дата и время
Дополнение данных
Замена
Калькулятор
Кросс-таблица
Объединение
Параметры полей
Разгруппировка
Свертка столбцов
Скользящее окно
Слияние
Соединение
Сортировка
Фильтр строк
Управление
Выполнение узла
Подмодель
Узел-ссылка
Условие
Цикл
Исследование
Автокорреляция
Качество данных
Корреляционный анализ
Факторный анализ
Предобработка
Заполнение пропусков
Квантование
Конечные классы
Разбиение на множества
Производные компоненты
Подключения
SQLite (CRM)
Телефоны (сайт)
Телефоны (маркетинг)
Телефоны (CRM)
Объединение всех данных
Удаление дублей
Очистка телефона
Войти
SMS
Список: SMS-рассылка
Мобильные телефоны
Список: колл-центр
Оценка качества
Распределение потока
Список: ручной разбор
Реальный сценарий:
Подготовка данных
Очистка
Постобработка

Телефон – пример очистки

ab ID	8
ab Исходный номер телефона	8 908 92 24933 справочная
ab Очищенный номер телеф...	+7 (908) 9224933
ab Добавочный номер теле...	
ab Тип номера телефона	Мобильный
ab Остаток строки	справочная
ab Код страны	7
ab Код ABC/DEF	908
ab Номер телефона	9224933
ab Страна	Россия
ab Регион	Свердловская область
ab Город	
ab Оператор	ООО "ЕКАТЕРИНБУРГ-2000"
ab Часовая зона 1	МСК+2
ab Часовая зона 2	московское время плюс 2 часа
ab Часовая зона 3	UTC+5
ab Лог	В номере 11 цифр и первая 8, она заменяется на 7
7 Дата очистки	20.10.2017 18:00:20

Исходные данные

Результат обработки

Почтовый адрес – пример очистки

ab ID	252162
ab Исходный адрес	ЭНГЕЛЬС-23,УЛ.,АТКАРСКАЯ,24
ab Найденный адрес	413110, Саратовская область город Энгельс улица Аткарская д. 24
9.0 Доля найденных букв запроса	0,783
12 Масштаб адреса	4
ab Код КЛАДР	6400000130000006
ab Ключ ФИАС	cd6746ba-1dab-4dc7-a86e-c84b71fdaf5b
ab Почтовый индекс	413110
0/1 Признак актуальности	Ложь
ab Полный адрес актуального АЭ	Саратовская Область Энгельсский Район Рабочий поселок Приволжский Улица Аткарская
ab Код КЛАДР актуального АЭ	640390000770037
ab Ключ ФИАС актуального АЭ	6272133e-3bb0-4dff-acee-8e0293387f20
ab Почтовый индекс актуально…	413110
ab Тип субъекта РФ	Область
ab Наименование субъекта РФ	Саратовская
ab Тип "города"	Город
ab Наименование "города"	Энгельс
ab Тип "улицы"	Улица
ab Наименование "улицы"	Аткарская

Исходные данные

Результат обработки

Почтовый адрес – результат очистки

3 группы адресов по качеству:

- Белая, можно использовать – адрес подтвержден и разобран до дома
- Серая, на ручную проверку – не найден дом, есть коллизии, нет уверенности
- Черная, не пригоден для работы – не почтовый адрес, невозможно идентифицировать адрес

Сырой адрес	Результат очистки	Зона
СОЧИ, ул. ШОССЕЙНАЯ д. 5В кв. 25, 354037	354037, Краснодарский край город Сочи улица Шоссейная д. 5В, кв. 25	Белая
614000;Пермский край;Город ПЕРМЬ;Улица СЕМФИРОПОЛЬСКАЯ;д. 86	614089, Пермский край город Пермь улица Симферопольская д. 86	
644117 обл Омская Г. ОМСК УЛ. 3МОЛОДЕЖНАЯ дом 62 корп. СЕК. 5 кв. 64	644117, Омская область город Омск улица Молодежная 3-я д. 62, кв. 64	
414009 Астрахань Атарбекова 17 12	414009, Астраханская область город Астрахань улица Атарбекова д. 17, кв. 12	Серая
	414009, Астраханская область город Астрахань площадь Атарбекова д. 17, кв. 12	
	414009, Астраханская область город Астрахань переулок Атарбекова д. 17, кв. 12	
край Краснодарский Г. КРАСНОДАР ПР-КТ. И-М. ПИСАТЕЛЯЗНАМЕНСКОГО дом 14 кв. 157	350065, Краснодарский край город Краснодар проспект им писателя Знаменского д. 14, кв. 157	
Самара в/ч 12454	443000, Самарская область город Самара в/ч 12454	
Калужская область Д ПЕРЕДОЛЬ улица НЕТ стр. д. 10 кв.	249160, Калужская область Жуковский район деревня Передоль д. 10	
2 а ул. Усть-Курдюмская	410018, Саратовская область город Саратов улица Усть-Курдюмская д. 2А	
10А ул.м-он Солопова		Черная
10-Я КРАСНОАРМЕЙСКАЯ,22А		
Всеволожск нет Промзона Кирпичный завод		

Домен	Записей/сек.
Фамилия, имя, отчество	540
Название организации	262
Почтовый адрес	5
Телефоны	714
Электронная почта	9 091
Удостоверения личности	1 785
ИНН	2 272
Даты	7 142
Банковские реквизиты	562

Набор:

100 000 записей

Режим:

Пакетная обработка

Сервер:

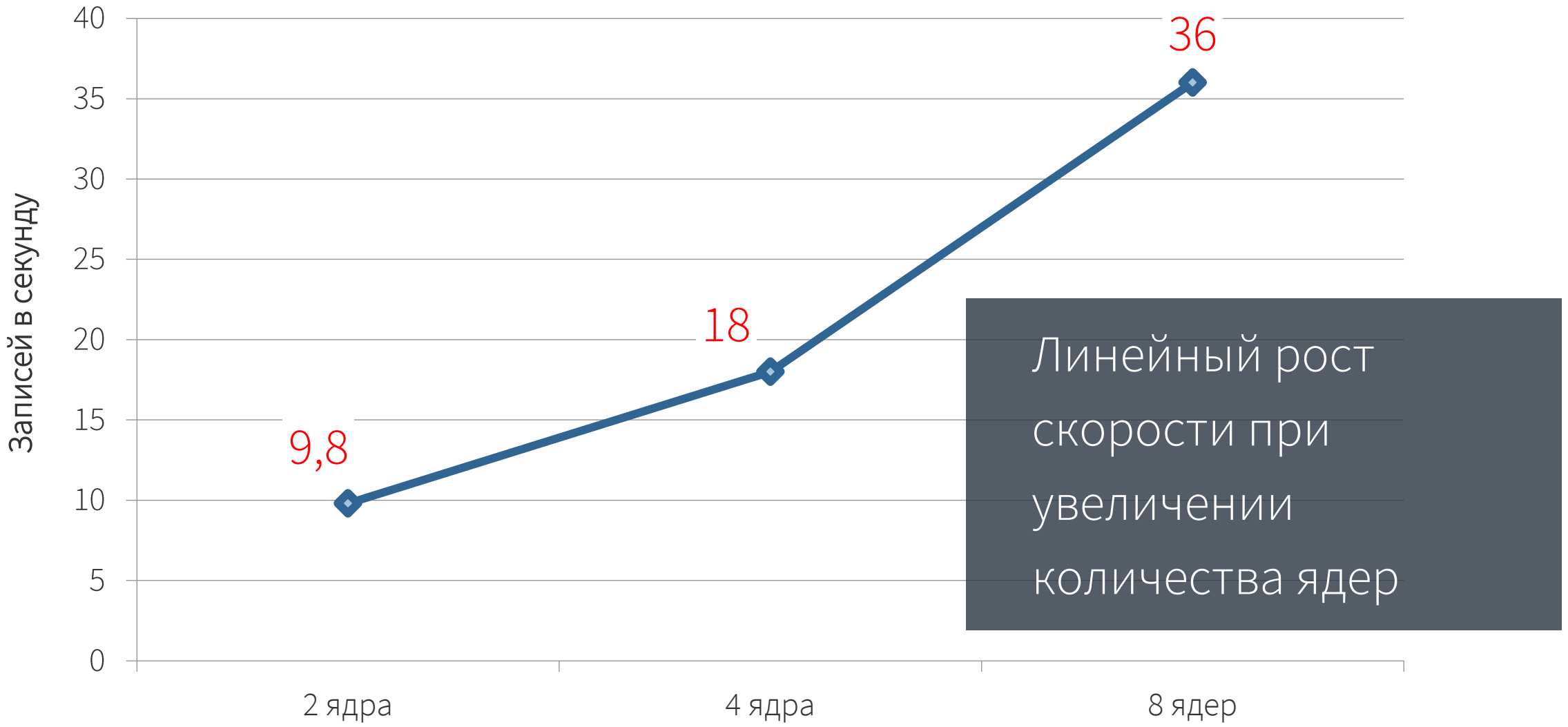
Windows Server 2012 R2 x64/ Intel Xeon E5, 4 ядра 2.60 GHz/16 Gb/1 Tb

Поставляемые справочники	Записей
Почтовые адреса: до улиц	1 909 597
Почтовые адреса: дома	32 592 474
Почтовые адреса: квартиры	54 280 979
Фамилии	752 367
Имена	10 088
Отчества	28 305
Операторы связи/регионы	396 037
Недействительные паспорта	128 631 627
E-mail домены	32 440

8

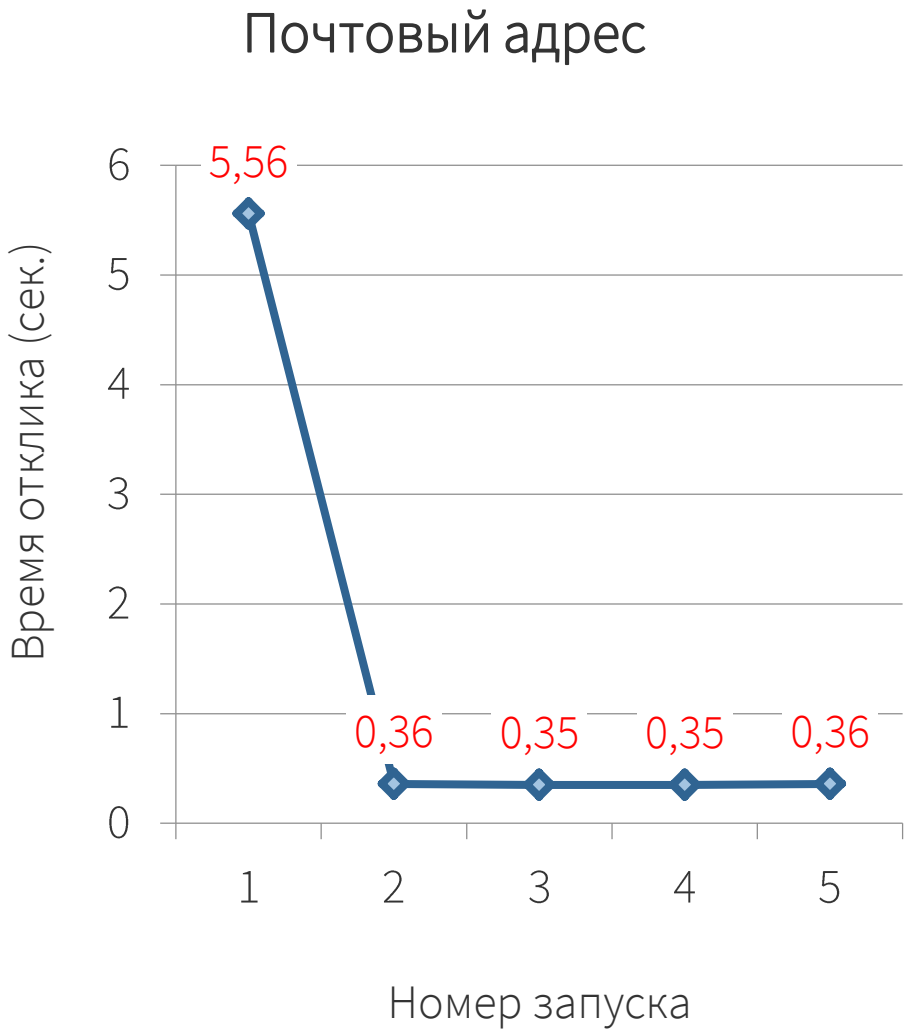
поставляемых справочников

Пример: Очистка почтовых адресов

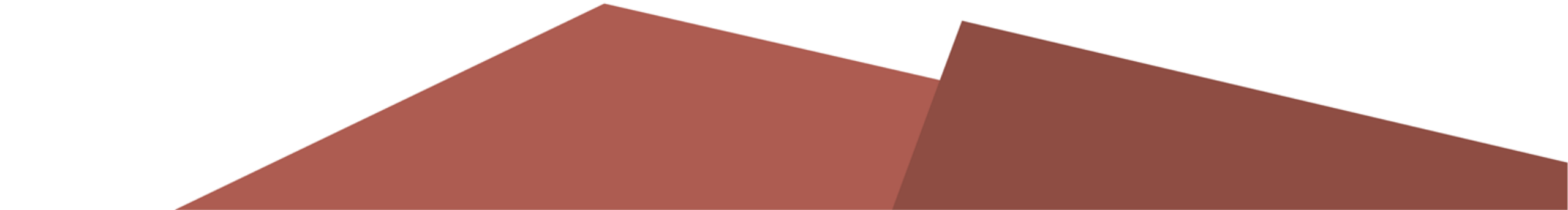


Веб-сервис:

1. Работает медленнее из-за накладных расходов
2. Сильнее зависимость от сетевой инфраструктуры
3. Нужно время на первое обращение – «разогрев»



Дедупликация и создание «золотой» записи



Конечная цель очистки клиентских данных – создание «золотой записи»: единого, точного, актуального и полного варианта записи клиента.

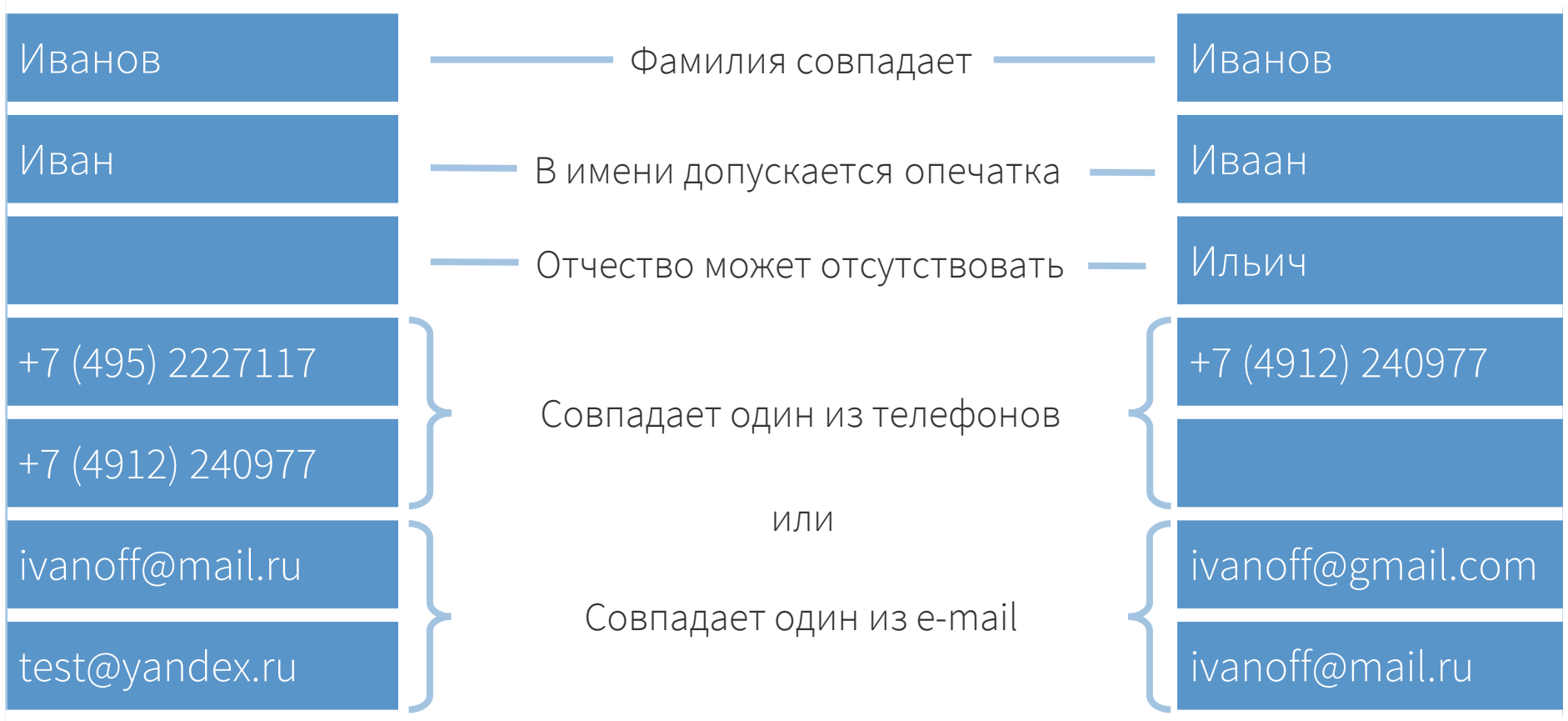
Для этого требуется:

1. Найти дубли и похожие записи
2. Объединить данные из всех источников

Возможности дедупликации:

1. Формирование групп дублей:
   - Задание сочетаний атрибутов, определяющих дубль
   - Нечеткий поиск с заданной точностью
   - Учет наличия пропусков
2. Создание эталонных записей
3. Формирование «серой зоны» групп дублей для ручной обработки

Пример стратегии дедупликации



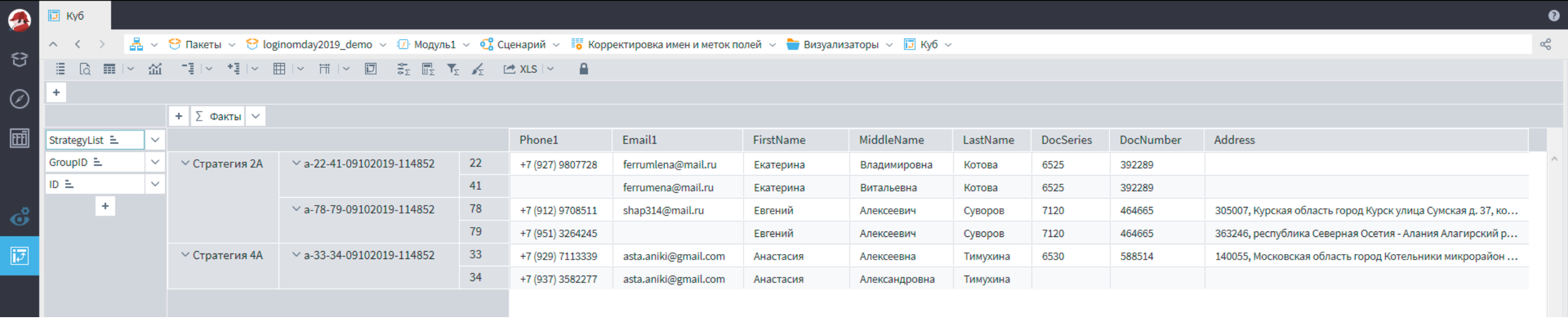
Преднастроенные стратегии дедупликации

Физические лица:

- 5 автоматических (жестких) стратегий
- 3 ручных (мягких)

Юридические лица:

- 5 автоматических (жестких) стратегий
- 3 ручных (мягких)



Пример стратегии дедупликация по жестким стратегиям для автоматической обработки

Атрибуты	Расстояние Дамерау-Левенштейна	Непротиворечивость
Страна	0	True
Имя	2/25	False
Фамилия	2/25	False
Инициалы	0	False
Тип документа	0	False
Серия документа	0	False
Номер документа	0	False

После дедупликации возможно формирование эталонной «золотой записи»* с учетом:

- Доверия к источнику данных
- Качества очистки
- Актуальности данных
- Заданной метрики

* В рамках проекта

Группа дублей

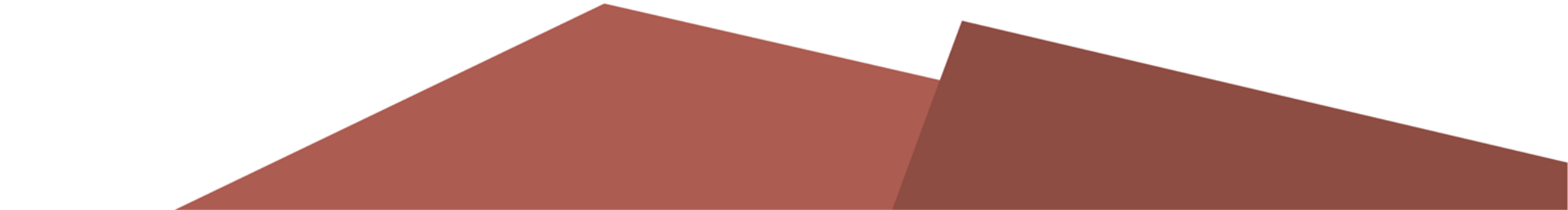
Фамилия	Имя	Отчество	Дата рождения	Телефон	Телефон	E-mail	E-mail
Иванов	Иван			+7 (495) 2227117	+7 (4912) 240977	ivanoff@mail.ru	test@yandex.ru
~~Иванов~~	~~Иваан~~	Ильич		~~+7 (4912) 240977~~		iii@yandex.ru	~~ivanoff@mail.ru~~
~~Иванов~~			15.12.1971	+7 (4912) 240699		~~ivanoff@mail.ru~~	

«Золотая» запись

Единичные атрибуты	Значение
Фамилия	Иванов
Имя	Иван
Отчество	Ильич
Дата рождения	15.12.1971

Множественные атрибуты	Значение
Телефон 1	+7 (495) 2227117
Телефон 2	+7 (4912) 240977
Телефон 3	+7 (4912) 240699
E-mail 1	ivanoff@mail.ru
E-mail 2	test@yandex.ru
E-mail 3	iii@yandex.ru

Поставка решения



Модуль решения	Компонент Loginom	Веб-сервис
Очистка ФИО	●	●
Очистка названия организации	●	●
Очистка почтового адреса	●	●
Очистка телефонов	●	●
Очистка электронной почты	●	●
Очистка удостоверений личности	●	●
Очистка реквизитов организаций	●	●
Очистка дат	●	●
Очистка банковских реквизитов	●	●
Дедупликация	●	
Формирование «золотой записи»	●	

Режим	Назначение
Пакетная обработка	1. Первичная очистка данных 2. Регулярная регламентная очистка 3. Эпизодическая очистка больших наборов данных
Веб-сервис	1. Online очистка 2. Очистка небольших наборов данных 3. Интеграция с другими системами

Техническая поддержка:

1. Ежемесячное обновление и расширение справочников
2. Регулярные улучшения логики очистки данных

Выгоды решения

1. **Быстрый запуск** – минимальное время от получения данных до первого результата
2. **Гарантии качества** – исправление сотен ошибок, актуальные справочники
3. **Визуальный конструктор** – интеграция и настройка логики без кодирования

loginom.ru