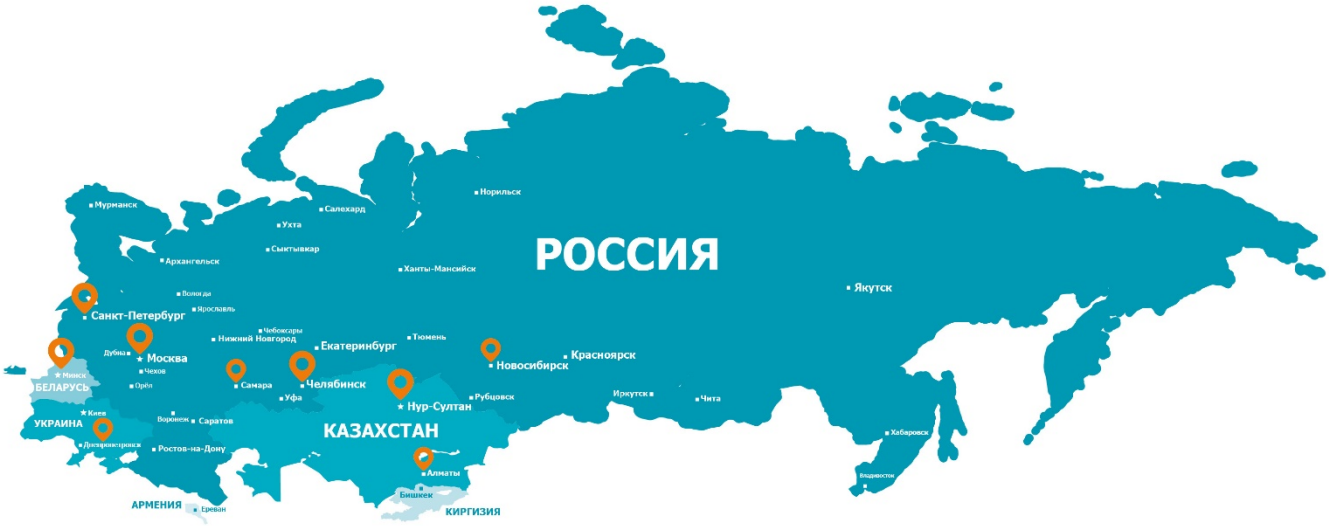
INVITRO

Данные тоже должны быть чистыми

Анастасия Рязанова,
ведущий аналитик-программист

Москва, 2019

Частная медицинская компания в России, специализирующаяся на лабораторной диагностике и оказании медицинских услуг. Представлена в 6 странах



ИНВИТРО – это

- 12,5+ млн пациентов в год
- 1300+ офисов
- 500+ городов присутствия
- 800+ серверов задействованы в работе
- 9000+ сотрудников в компании

Проблема

- Огромный объем поступающей информации
  - Общий размер клиентской базы 32 млн. записей (каждый день изменяются 64 тыс.)
- Разный уровень грамотности у администраторов медицинских офисов
- Некорректные персональные данные = недостоверные результаты
- Некорректные контактные данные = отсутствие информации о последних «вкусных» акциях и нововведениях

Проблема. Много точек входа при работе с клиентскими данными

Наполнение клиентской базы:

- администраторы в медицинских офисах
- операторы контакт-центра
- интеграционные автоматизированные сервисы клиентов
- сами пациенты (предзаказы в личном кабинете на сайте)

Обработка клиентской базы:

- операторы контакт-центра (при звонке на горячую линию)
- операторы баз данных (регулярно проводят анализ данных)

Проблема. Примеры «грязных» данных

Ошибки в ФИО

- Мари**а** (вместо Мария)
- **С**ветлана (первый символ был записан латиницей)
- **Саша** (вместо Александр)

Некорректно указан пол пациента или не указан совсем

Неполный номер телефона

- 454545 (без указания кода страны и города)
- +7 (только код страны)

Некорректный формат email

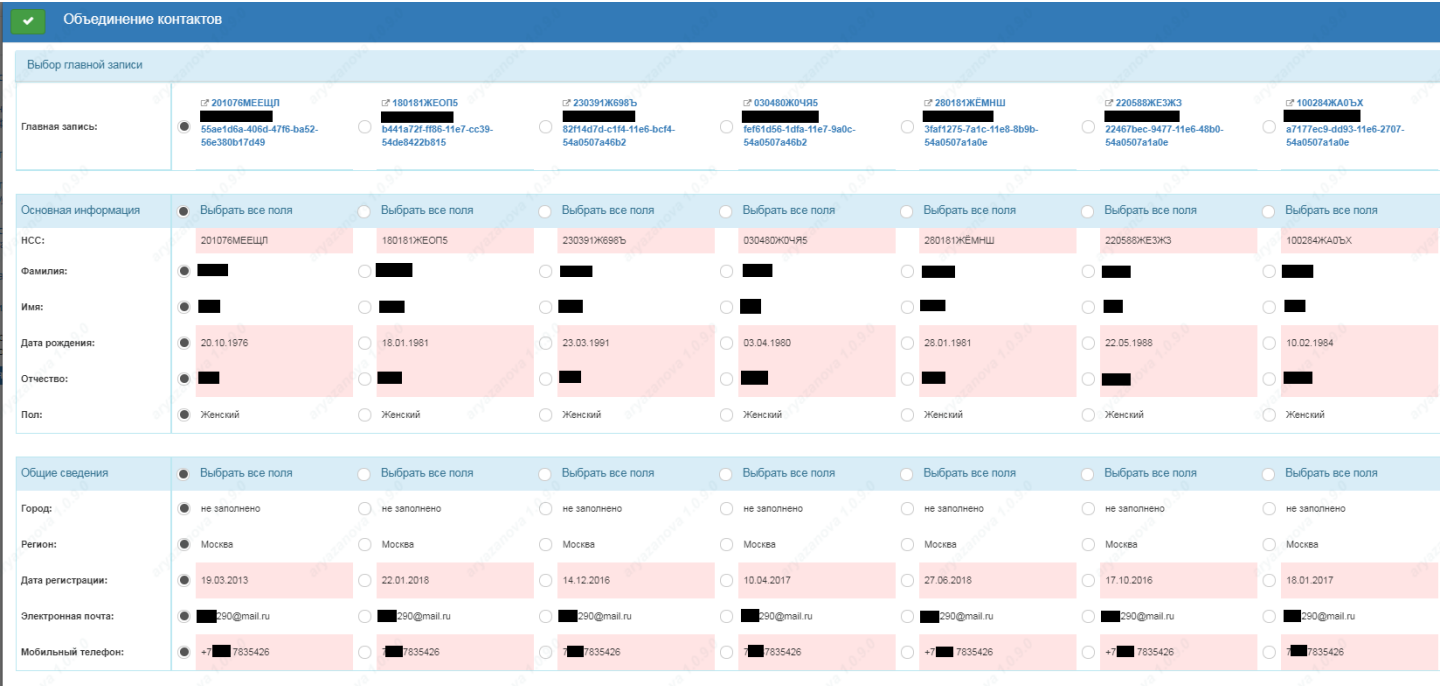
- test**.**mail.ru (test@mail.ru)
- test@gmail.**ru** (test@gmail.com)

Проблема. Дубликаты

Источники дублей:

- Особенности корпоративных правил и защиты персональной информации
  - администраторы медицинских офисов видят только ограниченную зону клиентов по региону размещения, а не всю базу
- Ошибки в персональных данных существующего в базе пациента
- Невнимательность администраторов при поиске пациента в базе

Проблема. Дубликаты. Пример



Группа из 9 дублей

Проблема. Дубликаты. Пример

Общие сведения	Выбрать все поля	Выбрать все поля	Выбрать все поля
Город:	не заполнено	не заполнено	не заполнено
Регион:	Москва	Москва	Москва
Дата регистрации:	19.03.2013	22.01.2018	14.12.2016
Электронная почта:	[redacted]290@mail.ru	[redacted]290@mail.ru	[redacted]290@mail.ru
Мобильный телефон:	+7[redacted]7835426	7[redacted]7835426	7[redacted]7835426

В группе полностью идентичны фамилия, пол, номер мобильного телефона, адрес эл. почты.

Незначительные различия в именах и дате рождения.

Задачи

- Наладить динамический процесс проверки информации, допускающий внесение изменений в функционал сервиса
- Снизить влияние человеческого фактора при вводе и обработке контактных данных
- Автоматизировать обработку, очистку и дедупликацию
- Поддерживать гарантированное качество персональных данных клиентов
- Учитывать специфику бизнес-процессов компании

Решение

В 2013 году была создана система по очистке и дедупликации данных на базе аналитической платформы Deductor. В 2019 решение было перезапущено на базе новой версии – Loginom.

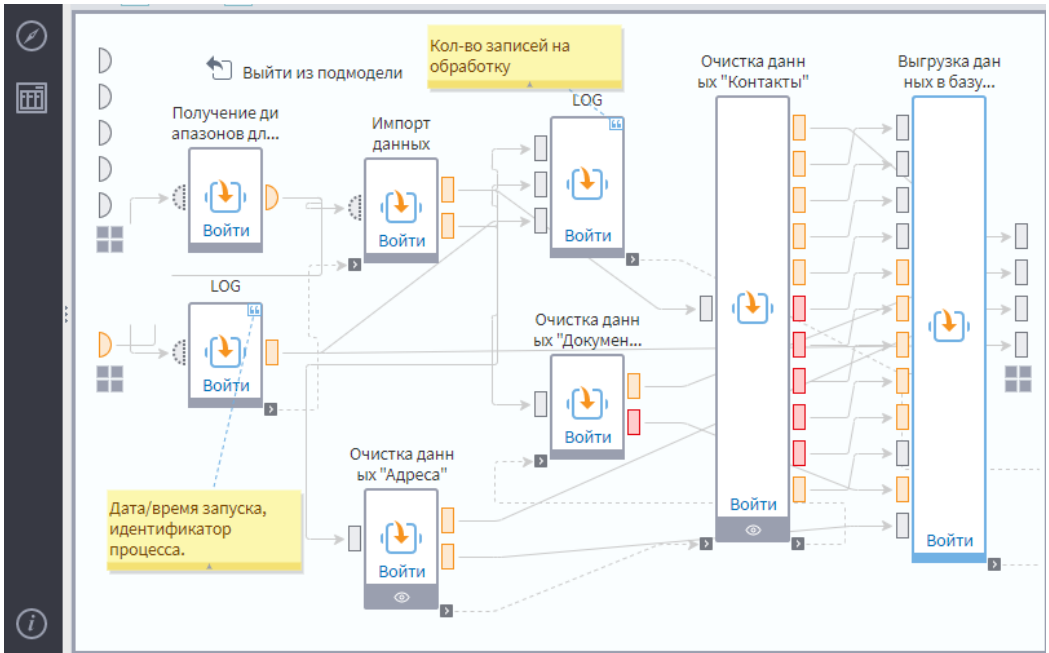
Предпосылки перехода на новую версию

- Заканчивалось технологическое развитие старой системы, а бизнес-процессы требовали изменений
- Система уже не справлялась с объемом измененных данных и просто «захлебывалась» в их количестве

О системе очистки и дедупликации

Loginom

Loginom – очень гибкая система, которая позволила нам максимально «подстроить» ее под наши технологические и бизнес-процессы.



Состав системы

В основе системы использовано решение Loginom Data Quality

- Очистка
  - Исправление ошибок, проверка по справочникам, стандартизация, обогащение
- Дедупликация
  - Проверка дублей по стратегиям
  - Объединение дублей в группы, автоматическое слияние или передача операторам для уточнения информации и принятия решения о слиянии

Входные данные

Система работает с несколькими справочниками:

- Пациенты (ФИО, пол, дата рождения, контактные данные)
- Адреса пациентов
- Документы пациентов

Очистка и обогащение на примере ФИО

- Удаление лишних символов в ФИО (**,** ; **"**; **–**)
- Проверка основного языка написание ФИО, замена символов другого языка аналогами основного (латинская **C** вместо кириллической)
- Сверка имени, отчества с постоянно обновляемыми справочниками имен, отчеств, фамилий, встроенных в Loginom
  - Исправление орфографических ошибок
  - Восстановление до полного имени
- Восстановление пола по ФИО в случае его отсутствия

Дедупликация

- Только уже очищенные данные
- Поиск дублей по полной базе пациентов по заложенным 6 четким и 3 нечетким стратегиям
- Объединение в группы полных и потенциальных дублей
- Передача групп потенциальных дублей операторам
- Выделение «золотой» мастер-записи в группах полных дублей
- Слияние дублей в группе на мастер-запись
- Публикация изменения в шину данных

Дедупликация. Пример стратегии

СТРАНА (непротиворечивость=true, точное совпадение) И

ИМЕНА (непротиворечивость=false, нечеткость 2/20) И

ФАМИЛИИ (непротиворечивость=false, нечеткость 2/20) И

ИНИЦИАЛЫ ОТЧЕСТВА (непротиворечивость=false, точное совпадение) И

ТИП ДОКУМЕНТА (непротиворечивость=false, точное совпадение) И

СЕРИЯ ДОКУМЕНТА (непротиворечивость=false, точное совпадение) И

НОМЕР ДОКУМЕНТА (непротиворечивость=false, точное совпадение)

Дедупликация. Пример стратегии

Нечеткие совпадения базируются на расчете расстояния редактирования Дамерау-Левенштейна (РР)

Нечеткость **2/20** означает, что допускается РР не больше **2 символов** и процент РР от общей длины не больше **20%**

НАТАЛ**Ь**Я = НАТАЛ**И**Я (разница в 1 символ и 14% от общей длины) Нечеткость 1/14 – будут признаны идентичны.

НАТАЛ**ЬЯ** = НАТАЛ**И** (разница в 2 символ и 28% от общей длины) Нечеткость 2/**28** – **не будут** признаны идентичны

Процесс обработки данных. Шаг 1

Петров
Ваня
Г
01.01.1986
Муж
89111111111

1 визит: 09.10.2017

Петров
Иван
Геннадьевич
02.01.1986
Жен
79111111111

1 визит: 01.02.2019

Иванов
Иван
Генадьевич
01.01.1986
Муж
79111111111

1 визит: 14.04.2019

Очистка данных

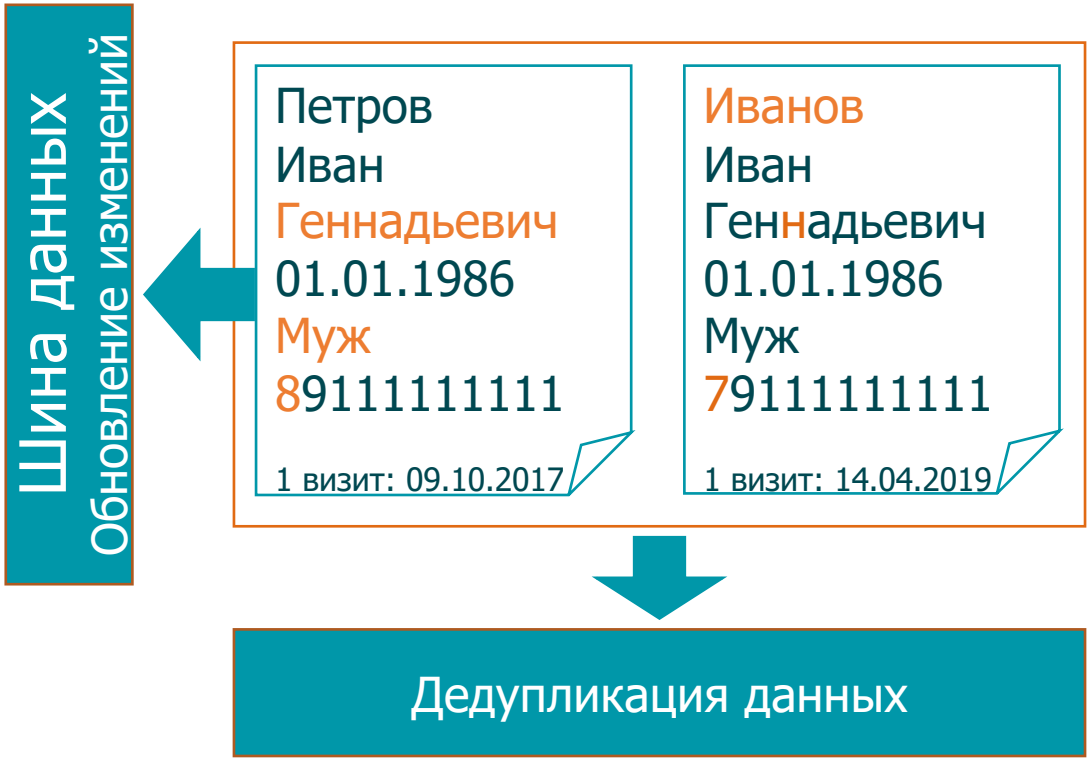
Процесс обработки данных. Шаг 2

Петров Иван Г 01.01.1986 Муж 89111111111 1 визит: 09.10.2017	Петров Иван Геннадьевич 02.01.1986 Муж 79111111111 1 визит: 01.02.2019	Иванов Иван Геннадьевич 01.01.1986 Муж 79111111111 1 визит: 14.04.2019

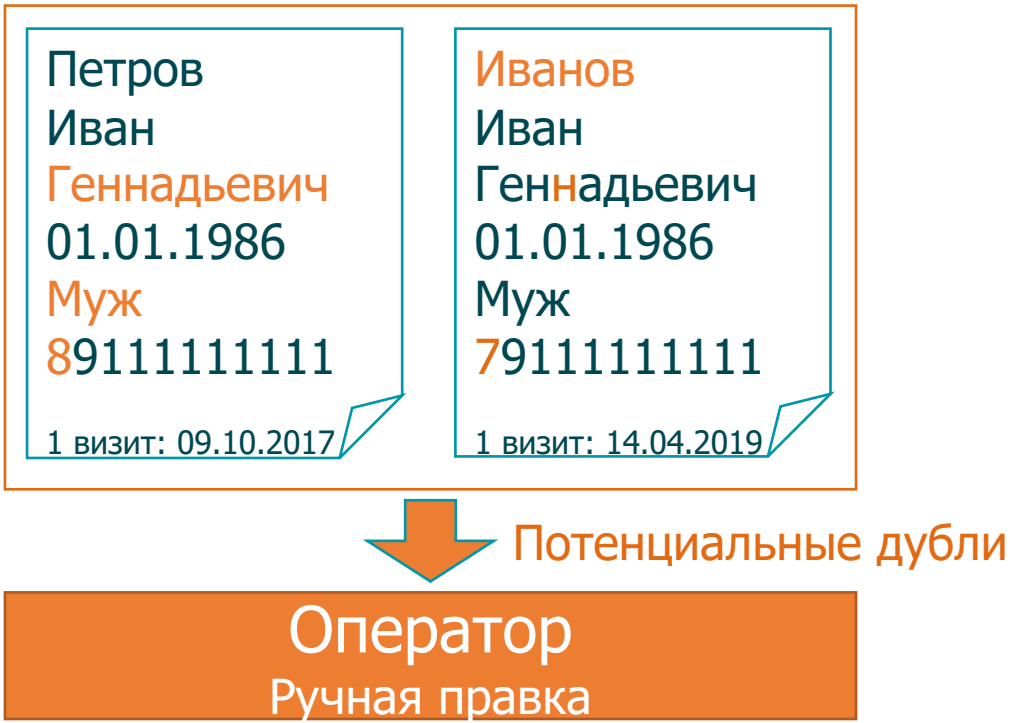
Полные дубли

Дедупликация данных

Процесс обработки данных. Шаг 3



Процесс обработки данных. Шаг 4



Взаимодействие систем

Два направления взаимодействия:

Частичная репликация клиентской базы и базы Loginom

- Для обработки выбираются все измененные за **5 минут** записи
- Записывает потенциальные дубли в отдельную таблицу, доступную только оператору

Шина данных

- После очистки и дедупликации Loginom публикует изменения в шину данных
- Остальные сервисы, работающие с клиентской базой, «разбирают» изменения из шины

Статистика по работе

- Ежедневно
  - Очищаются около 50 тыс. записей
  - Обогащаются около 15 тыс. записей
  - Сливаются около 4 тыс. дублей
  - Ручная обработка оператором 50-100 записей

- За все время работы сервиса
  - Очищены более 5,5 млн записей
  - Проверены на дубли около 3 млн записей

Состав команды ИНВИТРО

В процессе внедрения Loginom участвовали:

- 1 сотрудник группы поддержки лабораторных систем
- 1 сотрудник отдела разработки
- 1 архитектор баз данных

Перспективы в развитии системы

- Подключить к системе другие справочники
- Использовать более широкий спектр возможностей Loginom в глубокой аналитике данных

Результаты

- Результативность внедренной системы подтвердила планы
  - Снижены издержки на поддержку качества клиентских данных
  - Существенно уменьшен процент попадания некорректных и неполных данных в клиентскую базу
  - Успешно запущен процесс «избавления» от дублей

INVITRO

Спасибо за внимание!

www.invitro.ru

8 (800) 200-363-0