H F Labs

Февраль, 2023

Маскировщик.

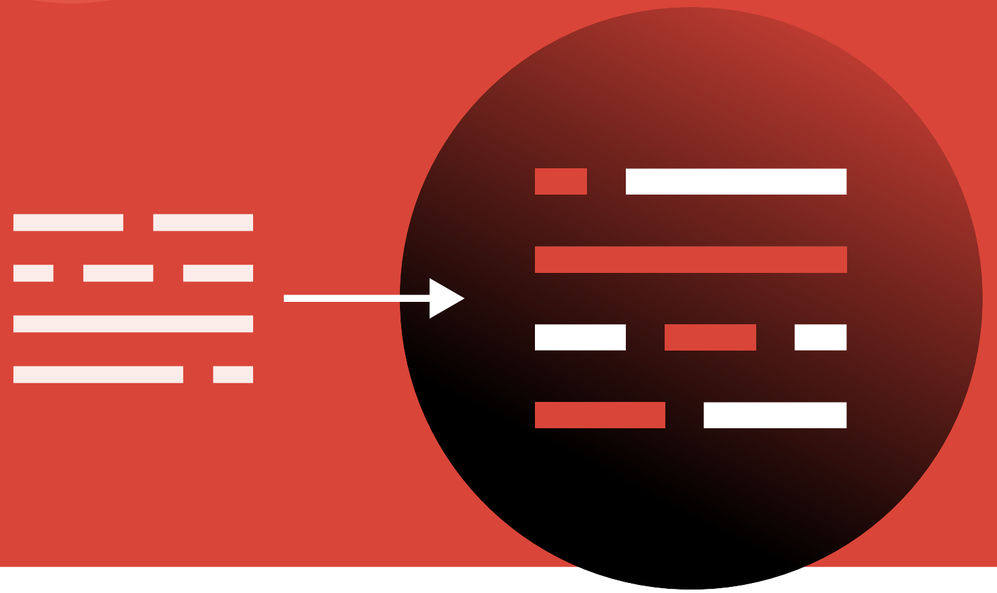
Маскируем данные без потери смысла

Михаил Берёзин, CPO

mikeberezin.t.me

Проблема:

Зачем нужно маскирование?



Компании хотят избежать утечек персональных данных клиентов. Поэтому ограничивают доступ сотрудников к персданным.

H F Labs

Типовые задачи

Доступ к данным

Доступ к данным на боевых средах находится под строгим контролем, тогда как доступ к тестовым средам открыт большему количеству людей.

Для тестирования

Реальные системы требуют тщательного тестирования. Чтобы делать его качественно, необходимо проверять на очень похожих данных.

Построение моделей

Для построения аналитических моделей (в том числе, с привлечением подрядчиков), также необходим доступ к данным.

Есть данные в одной БД (источник), нужно сделать магию и положить их в другую БД (приемник). К ней уже имеют доступ те, кто не должен видеть оригинальные данные.

Но смысл их важен!

H F Labs

Подход 1

Замена на «Звёздочки»

Меняет тип данных

Поэтому алгоритм не годится для обезличивания чисел и дат

Слабо защищает популярные имена и адреса

Даже если сократить число «звездочек» до одной, популярные имена, отчества и города легко расшифровывать

Убивает смысл данных

Обезличенные данные теряют семантику, валидность, социально-демографические характеристики и связи по домохозяйствам

H F Labs

Подход 2

Замена букв на буквы, цифр на цифры

Поддается расшифровке

Если алгоритм замены простой, то обезличенные данные можно восстановить

Портит качество и полноту данных

Заменяя случайными буквами и цифрами осмысленные, мы теряем семантику, валидность, создем характеристики, связи по домохозяйствам.

Фамилия, Имя, Отчество Еременко Наталья Сергеевна	→	Нонингел Гпьпдбч Мношннагп
Дата рождения 21.07.1961	→	11.02.1973
Паспорт 6806 108711	→	7187 315818
Телефон 8 926 118 1212	→	9 037 229 2323
Адрес Тамбов, ул. Советская, 11	→	Тамбов

H F Labs

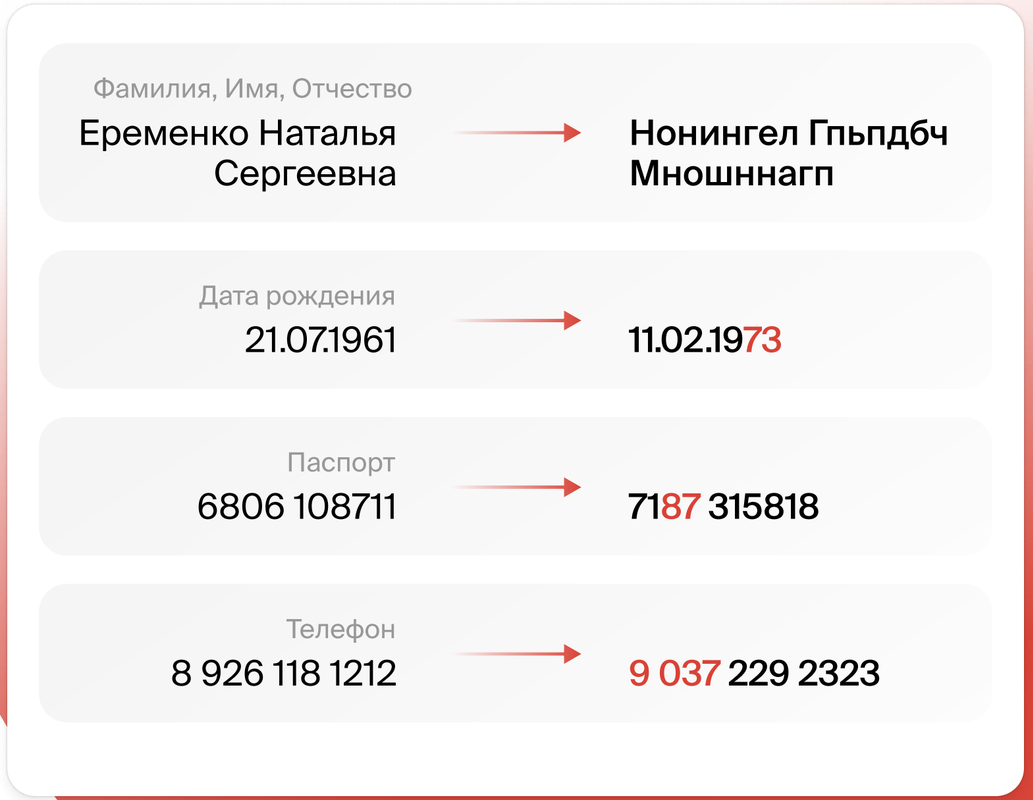
Подход 2: типичный кейс

Замена букв на буквы, цифр на цифры

При замене 61-летняя москвичка с действительным паспортом превратится в конструкт неизвестного пола в возрасте 49 лет, с несуществующим паспортом и номером телефона.

Утратили информацию про пол и испортили данные о возрастной группе, стране, регионе, паспорте и номере телефона. Потеряли возможные связи по ФИО и адресу с другими людьми в базе.

С такими данными невозможно провести соцдем-исследование или построить правдоподобную модель. Тестировщикам тоже придется трудно — данные не пройдут форматно-логических проверок.



H F Labs

Физлица

Как работает «Маскировщик»?

Еременко Петр Сергеевич

21 июля 1960

Ванина 1, Тамбов

6806 108771

8 926 118-12-12

mario@gmail.com

Сохранит родственные связи. Если Петра превратил в Сергея, то Петровичей — в Сергеевичей

Антонов Сергей Андреевич

11 февраля 1961

Дорожная 5, Тамбов

6807 203771 8 926 311-89-84 elf@mail.ru

Сохранит гендерный баланс

Сохранит социально-демографические группы

В адресе сохранит страну, город и район

Серию паспорта привяжет к году рождения, чтобы она прошла форматно-логический контроль

Оставит оператора и страну для номера телефона

Понимает домены емейлов: рабочие, личные, одноразовые такими и останутся

H F Labs

Домохозяйства

Как работает «Маскировщик»?

Иванов Петр Ильясович

21.07.1961

Тамбов,
Советская 11

Петров Андрей Фатихович

11.02.1961

Тамбов, Астраханская 5

Иванов Ильяс Петрович

05.03.1988

Тамбов,
Советская 11

Петров Фатих Андреевич

11.08.1987

Тамбов, Астраханская 5

«Маскировщик» понимает, что Ивановы Петр Ильясович и Ильяс Петрович, живущие по адресу город Тамбов, улица Советская, 11 — родственники. И учитывает эту информацию, когда маскирует их ФИО и адрес.

H F Labs

И что в итоге?

Из базы данных, содержащих записи о миллионах реальных людей, «Маскировщик» делает ровно такую же, неотличимую для человека базу данных людей, только нереальных.

Сохраняет родственные связи, качество данных, соцдем и географическое распределение

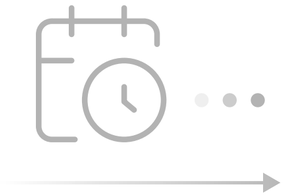
H F Labs

«Маскировщик» безопасен?

Да, абсолютно

11 НОЯБРЯ

Петр	→	Степан
Ирина	→	Татьяна
Наталья	→	Марина



15 ДЕКАБРЯ

Петр	→	Иван
Ирина	→	Марина
Наталья	→	Татьяна

В каждой сессии «Маскировщик» подбирает замены случайным образом.

Данные в файле замен хранятся в виде хеша, набора цифр и букв. Не зная алгоритма, восстановить исходные персональные данные невозможно. По умолчанию кэш сессии удаляется по её завершению.

H F Labs

Какие типы данных понимает «Маскировщик»?

Фамилии, имена и отчества

01.

Даты рождения

02.

Документы физлиц

03.

Адреса

04.

Номера телефонов

05.

Email адреса

06.

ИНН

07.

СНИЛС

08.

Номера банковских карт

09.

Номера банковских счетов

10.

Отличает типы данных и понимает их структуру. Качество данных не портится.

H F Labs

Почему «Маскировщик» это умеет?

18 лет

Опыт в качестве данных

СБЕР · РОСГОССТРАХ · Агентство по страхованию вкладов · госуслуги · OZON · Альфа-Банк · МегаФон

Все клиенты ↗

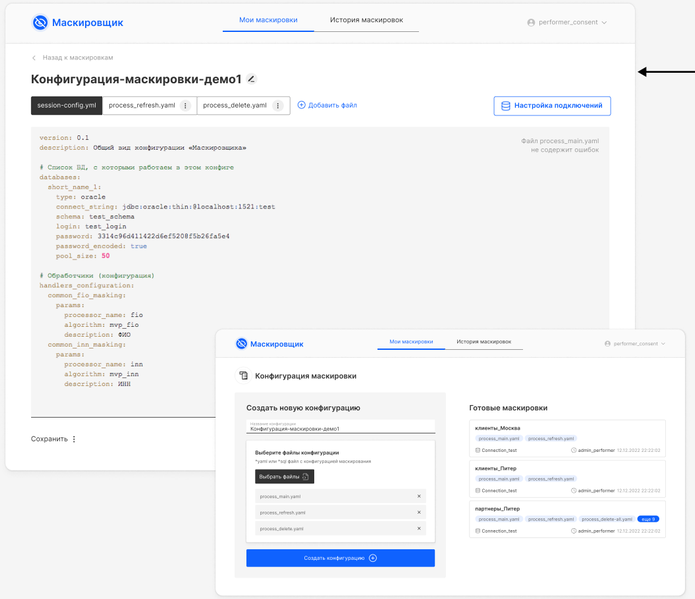
Раньше только стандартизировали, **теперь еще и маскируем!**

Под капотом те же уникальные алгоритмы

Перед маскированием данные разбираются из строки, понимаем, где и что должно быть

H F Labs

Как это работает технически?



Удобный веб интерфейс позволяет настраивать конфигурацию маскирования, а также смотреть историю сессий маскирования. Разграничен по правам доступа.

ETL сервис

чтение и запись данных, использование внешних сервисов для их обработки

чтение

запись

Источник

Приемник

базы данных на чтение и запись. Сейчас доступны Postgres и Oracle, но нет ограничений на подключение новых

PostgreSQL ORACLE

Набор сервисов маскирования

Умная замена данных

Кэш

Сохранение абсолютной консистентности в рамках сессии

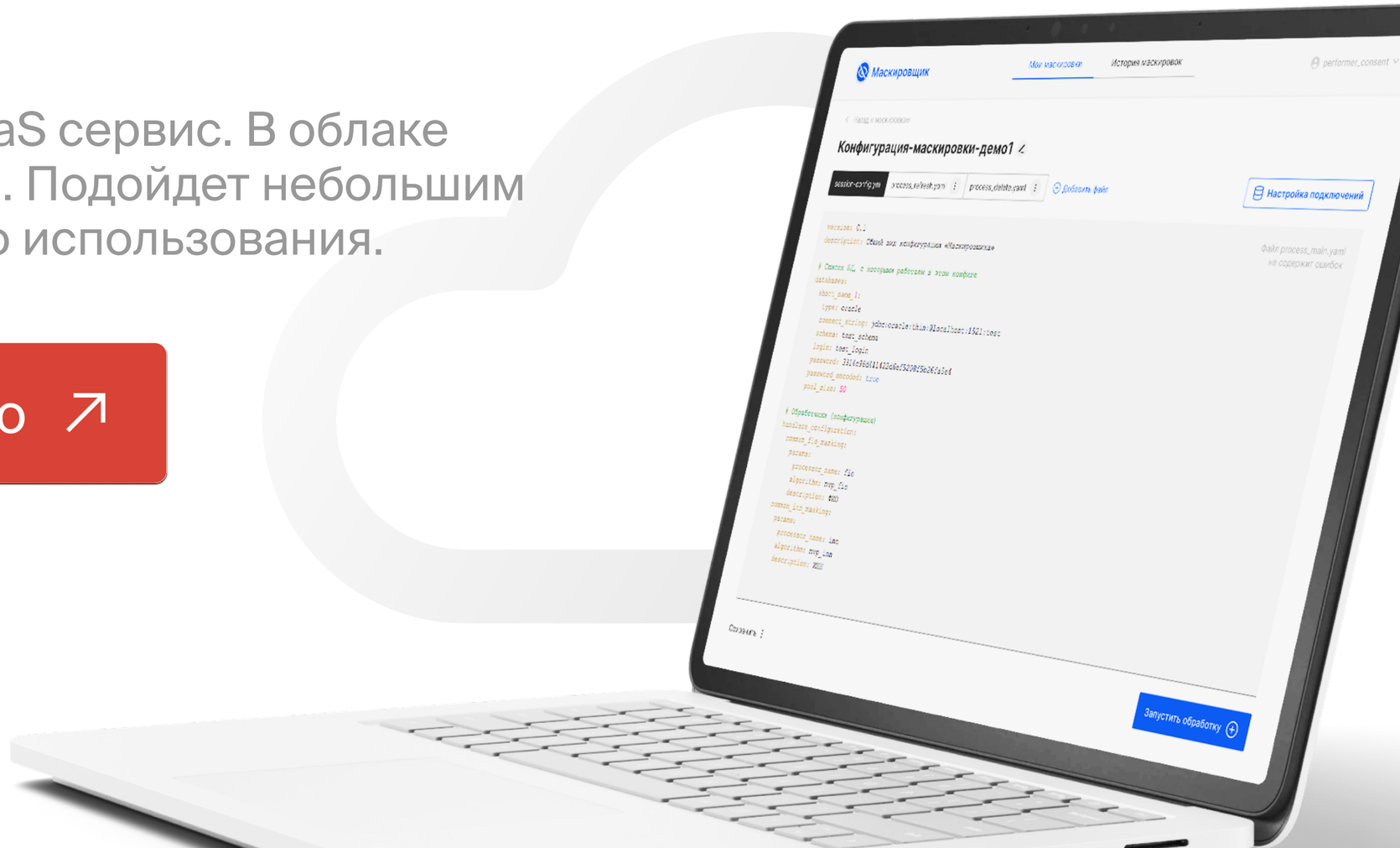
H F Labs

Еще и облачное решение DaData

По умолчанию «Маскировщик» — коробочное решение.

Но доступен так же и как SaaS сервис. В облаке доступны все те же сервисы. Подойдет небольшим компаниям или для частного использования.

Заявка на бета-версию ↗



H F Labs

H F Labs

Спасибо за внимание.

Задавайте вопросы

Страница про «Маскировщик»