

Система обезличивания данных для тестовых полигонов банка

Разработка программного обеспечения для финансовых организаций – это сложный и ответственный процесс, требующий соблюдения строгих стандартов безопасности и надежности. Важную роль при разработке играет этап тестирования функциональности продукта. Для эффективного тестирования и анализа необходимы большие массивы данных с минимальным отличием от реальных по объемам и качеству.

Банки обрабатывают огромное количество конфиденциальной информации, и ее защита – это одна из самых важных задач для финансовых организаций. Поэтому при разработке нового банковского программного продукта для проверки и тестирования его функциональности нельзя использовать имеющиеся данные клиентов без риска нарушения законодательства.

Для соблюдения правил и законов в области защиты данных применяют их обезличивание. Обезличивание данных для тестовых полигонов – это процесс удаления или замены идентифицирующей информации, такой как имена, адреса, номера телефонов и другие личные данные,

которые могут использоваться для идентификации конкретных лиц, из набора данных, используемых для тестирования или исследования.

В 2021 году ВТБ внедрил платформу обезличивания данных на базе алгоритмов машинного обучения. Это позволило повысить уровень защищенности персональной клиентской информации, а также банковской и коммерческой тайны.

Компания «Бизнес-Азимут» приняла участие в проекте по разработке системы создания и распространения обезличенных баз данных, используемых в процессе тестирования технологических продуктов. Продукт работает уже около 2-х лет и зарегистрирован в реестре российского ПО.

Система помогает защитить критически важную информацию пользователей и приложений, а также позволяет проводить автоматизированное обезличивание чувствительных данных с использованием методов машинного обучения и алгоритмов, сохраняя интеграционные связи, формат и смысл значений.

Обезличивание данных позволяет создавать аналитические и тестовые среды без раскрытия конфиденциальной информации клиентов. Банки могут использовать их для тестирования новых продуктов и услуг, анализа клиентского поведения, обучения персонала, а также для усовершенствования систем безопасности и предотвращения мошенничества.

Прежде всего, была создана модель профилирования баз данных, которая позволила определить список атрибутов персональных данных, то есть какая именно информация в базе относится к конфиденциальной и должна быть обезличена. В основе платформы используются Open Source решения в части ETL, UI, ML. Точность классификации данных составляет более 96%.

Важным критерием при использовании обезличенных данных при тестировании функциональности программного продукта является сохранение не только объема, но и формата данных. Для этого используется алгоритм шифрования с сохранением формата – FPE (Format-preserving encryption). Например, номер телефона должен остаться набором цифр идентичной размерности, а имя и фамилия должны быть переведены в данные с идентичным форматом и смыслом. Кроме того, критичным было сохранить прослеживаемость данных внутри и между системами. Например, один и тот же номер ИНН должен измениться на один и тот же обезличенный номер везде, где он фигурирует.

Функциональность системы:

- Создание обезличенных баз данных и обезличенных датасетов
- Создание облегченных баз данных с опцией настройки критериев усечения
- Загрузка данных из одной базы в другую без изменения
- Поиск и классификация в автоматическом режиме атрибутов с персональными данными на основе ML-моделей
- Разметка пустых полей и таблиц баз данных
- Использование библиотеки алгоритмов обезличивания, построенных на базе алгоритмов FPE-шифрования с сохранением формата значений и интеграционной целостности
- Формирование списков значений полей для исключения из обезличивания

Функциональность системы позволяет в кратчайшие сроки создавать еще более безопасные среды разработки и тестирования и сокращать время вывода продуктов и сервисов банка на рынок.

В настоящее время мы осуществляем следующий этап этого проекта - развитие и модернизацию системы, более точно затачивая ее под требования заказчика.

600
БАЗ ДАННЫХ
ОБЕЗЛИЧЕНО

2.5-4 ТБАЙТ
В СУТКИ
СКОРОСТЬ
ОБЕЗЛИЧИВАНИЯ

В 50 РАЗ
СОКРАТИЛИСЬ
ТРУДОЗАТРАТЫ НА
ОБЕЗЛИЧИВАНИЕ



БИЗНЕСАЗИМУТ