Преодоление барьеров:

Использование облачных сервисов и ИИ в условиях строгих ИБ-протоколов

Васильев Максим

2024

Васильев Максим: О спикере

Архитектура с 2017

Отчетность с 2009

@todmay

Почему это важно

- Отказ от технологий, таких как искусственный интеллект (ИИ) и облачные сервисы, может существенно ограничить конкурентоспособность бизнеса.

Почему это важно

- **Оптимизация и автоматизация бизнес-процессов:** ИИ и облачные технологии позволяют автоматизировать рутинные задачи, ускорять процессы и повышать точность решений.
- **Аналитика больших данных**: Позволяет компаниям собирать, хранить и анализировать большие объемы данных. Открывает новые возможности для улучшения клиентского обслуживания, прогнозирования рыночных тенденций и внедрения инноваций.
- **Гибкость и масштабируемость**: Облачные технологии предоставляют гибкость в управлении инфраструктурой, а ИИ помогает масштабировать аналитические процессы, обеспечивая более оперативное реагирование на изменения в бизнесе и внешней среде.

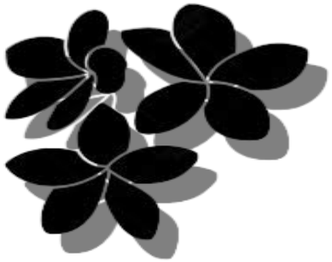
Преимущества

- **Экономия на инфраструктуре:** Облачные сервисы снижают необходимость в приобретении и обслуживании дорогостоящего серверного оборудования, предоставляя ресурсы по мере необходимости.
- **Ускорение процессов разработки и внедрения:** Внедрение ИИ и облаков ускоряет процессы разработки продуктов, улучшая время выхода на рынок и внедрение новых функций.
- **Увеличение производительности:** Благодаря ИИ бизнес получает доступ к инструментам, которые помогают повысить производительность сотрудников за счет более умного и автоматизированного анализа данных.

Аргументация запрета

- законодательная база и её трактовка
- неготовность принимать риски

Законодательная база



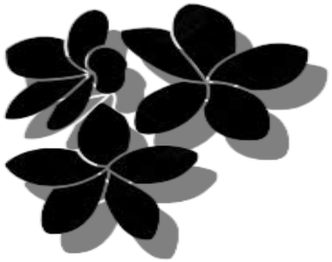
- регламенты и стандарты
- регуляторные нормы GDPR в Европе
  - Согласие на обработку данных.
  - Право на удаление данных ("право быть забытым").
  - Ограничение передачи данных в страны, где защита данных недостаточна.
- ФЗ-152 в России
  - сохранность данных в рамках юрисдикции РФ.

Неготовность принимать риски

- страх утечки данных
- нарушение регуляторных норм
- недостаток ресурсов для обеспечения необходимого уровня безопасности

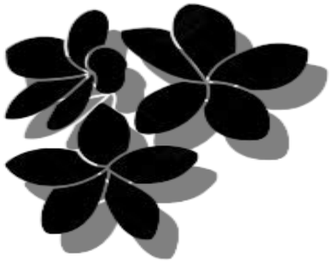
– часто это обусловлено недостатком знаний и опыта в этой области.

Топ 5 заблуждений о LLM



- Локальная версия LLM безопаснее
- Настроить безопасную среду для LLM в облаке невозможно
- Взаимодействие с LLM приводит к утечке
- LLM обучается на наших данных
- Другие клиенты увидят наши данные

Аргументы против мифов

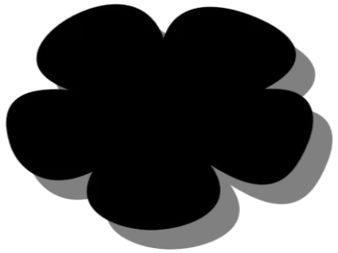


- Локальные инстансы требуют большего внимания
- Облака имеют стандарты по защите
- Запросы можно не логировать
- Инференс (обработка запроса) и обучение это разные процессы
- Архитектурная изоляция и возможность настройки гранулярных прав

Примеры из практики

- Деперсонализация данных
- Обезличивание данных
- Маркирование данных

Деперсонализация



Деперсонализация — это процесс удаления или замены персональных данных так, чтобы их нельзя было напрямую связать с конкретным человеком. Это позволяет использовать данные в аналитике и ИИ без нарушения законодательства о защите данных. Примеры: замена имен на уникальные идентификаторы или исключение чувствительных данных из наборов данных.

Как можно реализовать:

- Замена персональных данных идентификаторами
- Шифрование отдельных полей
- Псевдонимизация

Деперсонализация

Замена персональных данных идентификаторами: В базе данных можно заменить такие поля, как имя, адрес, или номер телефона, уникальными идентификаторами (например, UUID). Это позволяет сохранить структуру данных без персональной информации.

Шифрование отдельных полей: Для защиты персональных данных можно использовать симметричное или асимметричное шифрование для шифрования таких полей, как номера кредитных карт или социальные номера. Эти данные остаются полезными для обработки, но не могут быть легко расшифрованы без соответствующих ключей.

Псевдонимизация: В процессе псевдонимизации данные преобразуются так, что прямая идентификация пользователя невозможна без доступа к дополнительной информации. Например, можно сохранить ID клиента, но заменить его имя на псевдоним, а доступ к оригинальному имени ограничить.

PyCryptodome - библиотека шифрования для Python

Обезличивание данных

Обезличивание — это более глубокий процесс, при котором все данные, связанные с личностью, удаляются полностью. В результате никакая информация не может быть использована для восстановления личности даже косвенно. Это полезно при передаче данных третьим лицам или в публичные облачные сервисы.

Как можно реализовать:

- Удаление персональных полей
- Генерация синтетических данных
- Агрегация данных (объединение по группам)

Обезличивание данных

Удаление персональных полей: Перед отправкой данных в облако или ИИ-модель можно удалить все поля, содержащие персональную информацию (ФИО, адреса, контактные данные). Такие поля могут быть вырезаны или замещены нейтральными значениями.

Генерация синтетических данных: В некоторых случаях можно заменить реальные данные синтетическими, созданными на основе анализа исходных данных. Это позволяет сохранять схему и динамику данных без реальных персональных данных.

Агрегация данных: Объединение данных по группам так, чтобы не было возможности идентифицировать конкретного человека. Например, при анализе можно использовать средние значения по группам пользователей вместо индивидуальных данных.

Faker - библиотека синтетических данных для Python

Обезличивание данных

Согласно разъяснениям Роскомнадзора производные данные от персональных все еще остаются персональными, то есть хэширование или аналогичные алгоритмы не являются обезличиванием.

"В Роскомнадзоре уверены, что персональные данные, которые получены в результате обезличивания, остаются персональными. И формально, и юридически, потому что все еще характеризуют человека. А с дополнительной информацией и прямо на него указывают", - сказал он журналистам.

Это может быть важно когда ищем пересечения двух массивов данных.

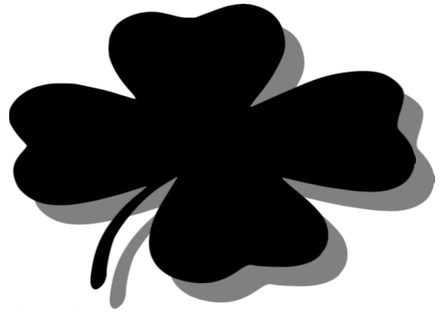
Маркирование данных

Маркирование — это процесс присвоения данным специальных меток или тегов, которые позволяют организовать информацию, отслеживать ее происхождение и контролировать доступ. Это важно для управления большими объемами данных, особенно в облаках, где метки могут помочь в идентификации чувствительных данных и обеспечении их безопасности.

Как можно реализовать:

- Добавление метаданных к каждому элементу данных
- Использование технологий контроля доступа (RBAC)
- Автоматизированные системы разметки данных

Маркирование данных



Добавление метаданных к каждому элементу данных: В процессе создания или обработки данных можно добавить специальные метки (теги), такие как уровень конфиденциальности или тип данных.

Использование технологий контроля доступа (RBAC): Метки данных можно связать с ролями доступа. Например, данные с меткой "секретно" могут быть доступны только определенным группам пользователей. Это реализуется через систему ролевого контроля доступа (Role-Based Access Control), которая ограничивает доступ в зависимости от меток.

Автоматизированные системы разметки данных: Существуют инструменты, такие как DLP-системы (Data Loss Prevention), которые автоматически помечают данные в зависимости от их содержания.

SpaCy – библиотека анализа естественного языка для Python

Что еще можно сделать

- Криптографические методы защиты
- Конфиденциальные вычисления (SMPC)

Что мы получим

- Возможность передавать данные без согласия
- Отвязку цифр от инсайда

Передавать без согласия

Использование методов деперсонализации и обезличивания позволяет передавать данные без необходимости получать согласие от субъектов данных, так как данные больше не идентифицируются как персональные. А также хранить данные после того как действие согласия закончилось.

Это критично для бизнеса, так как открывает доступ к масштабной обработке данных и аналитике без риска нарушений законодательства о защите данных.

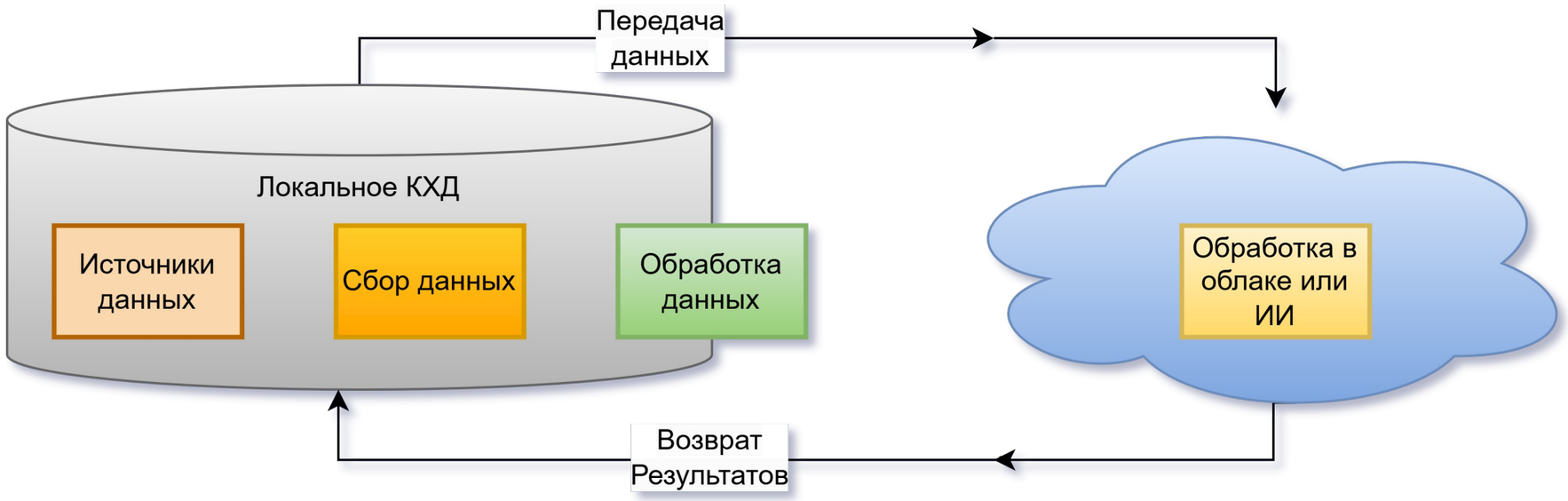
Отвязка цифр от инсайда

Отвязка цифр от инсайда означает использование обезличенных данных или псевдонимов, чтобы внутренние сотрудники и системы не могли сопоставить данные с реальными людьми. Это снижает риск утечек информации и минимизирует возможность инсайдерских угроз, улучшая безопасность внутри организации.

Архитектура решений

- Данные в облако
- Использование ИИ

Схема работы



Что дальше

- Уроки, извлеченные из реализации проектов
- Как изменение подходов к работе с данными повлияло на бизнес-процессы
- Повторное использование данных

Уроки

- **Безопасность данных и соответствие законам — приоритет.** Методы деперсонализации и обезличивания оказались критически важны для обеспечения безопасной передачи данных и соблюдения требований регуляторов, таких как GDPR и ФЗ-152.
- **Сложность балансировки между полезностью данных и их обезличиванием.** Чем глубже данные обезличиваются, тем меньше полезной аналитики можно извлечь. Важно найти баланс между сохранением полезных данных и их анонимностью.
- **Автоматизация процессов защиты данных.** Внедрение систем маркирования и шифрования данных требует автоматизации для эффективного масштабирования. Ручное управление метками не подходит для больших объемов данных.
- **Техническая готовность инфраструктуры.** Проекты показали необходимость заранее подготовленной инфраструктуры: шифрование, контроль доступа и аудит операций — обязательные элементы для успешного использования облаков и ИИ.
- **Упрощение взаимодействия с регуляторами.** Правильная деперсонализация и маркирование данных помогают компаниям избегать юридических проблем, позволяя безопасно работать с ИИ и облачными сервисами в условиях строгих ИБ-протоколов.

Изменения

- Ускорение обработки и анализа данных.
- Уменьшение бюрократии и административных барьеров.
- Повышение гибкости и масштабируемости.
- Снижение риска утечек и штрафов.
- Улучшение доверия партнеров и клиентов.

Повторение

- Эффективное повторное использование деперсонализированных данных. Данные, прошедшие деперсонализацию или обезличивание, можно использовать многократно без необходимости дополнительного согласования. Это значительно сокращает временные и административные затраты, позволяя применять их в различных проектах и аналитических задачах.

- Универсальность и гибкость данных. Один и тот же набор данных может быть повторно использован в различных бизнес-инициативах, включая обучение ИИ-моделей, маркетинговые исследования и аналитические отчеты, без нарушения законодательства или риска утечек.

- Создание многократной ценности из единого источника данных. Защищенные данные могут применяться в новых контекстах, обеспечивая бизнес новыми инсайтами и экономя ресурсы на сборе и подготовке данных.

- Минимизация затрат на управление данными. Повторное использование деперсонализированных данных снижает потребность в постоянной обработке новых наборов данных, что уменьшает расходы на хранение, защиту и обработку информации.

Основные выводы

Деперсонализация и обезличивание данных — ключ к безопасному использованию облачных сервисов и ИИ. Эти методы позволяют работать с данными без нарушения законодательных требований, ускоряя бизнес-процессы и снижая риски утечек.

Повторное использование данных значительно увеличивает ценность и эффективность работы с информацией. Обезличенные данные могут многократно применяться для различных целей, что снижает затраты и открывает новые возможности для аналитики и инноваций, при этом обеспечивая соответствие требованиям безопасности.

Написать мне:

@TODMAY

Материалы:

@SELFDATAARCH

Спасибо за внимание!