СБЕР
КИБЕР
БЕЗОПАСНОСТЬ

ISSN: 3034-2082

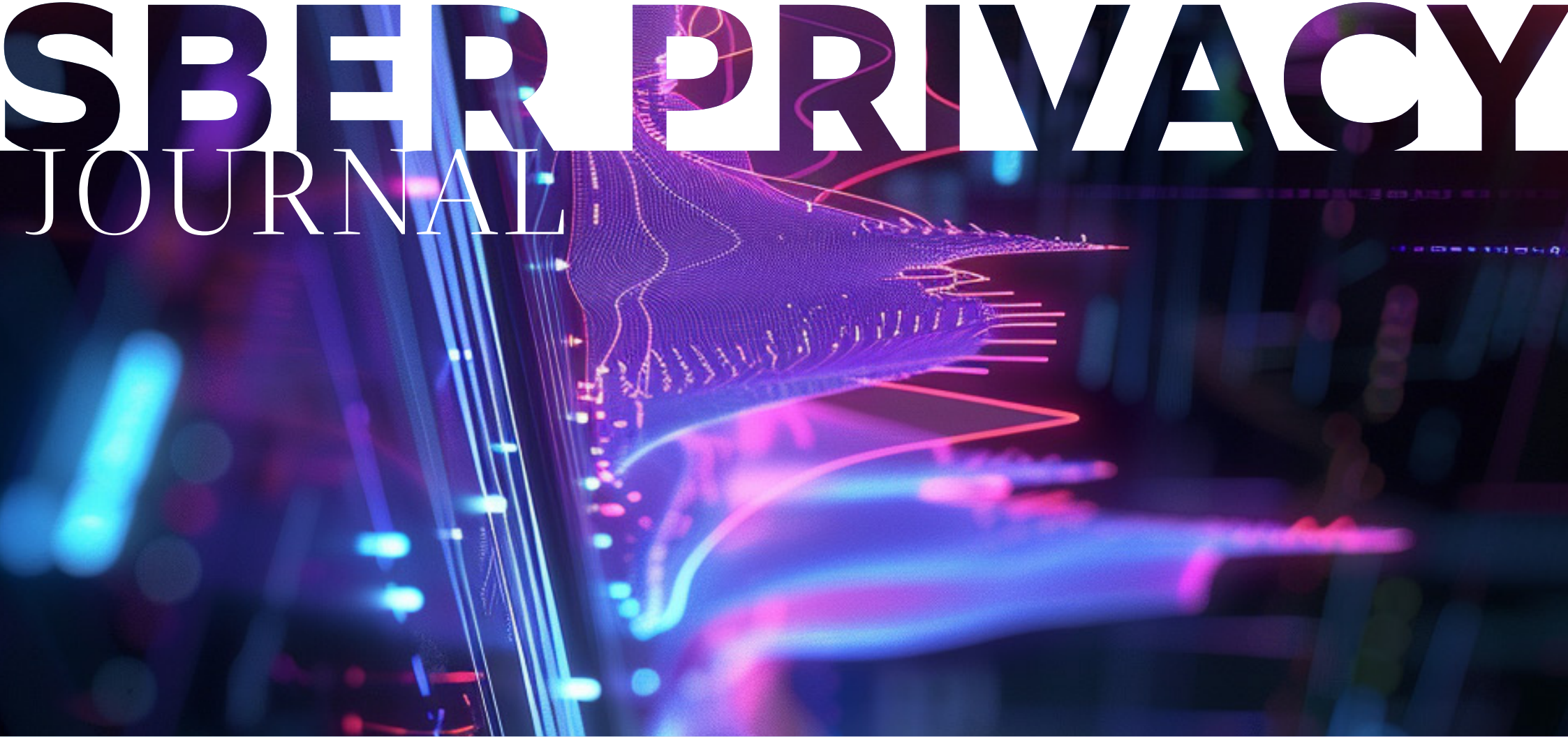
SBER PRIVACY JOURNAL

Журнал DPO о персональных данных
и приватности

ТЕМА НОМЕРА:
УПРАВЛЕНИЕ ДАННЫМИ В ЭПОХУ BIG DATA

ВЫПУСК №9 | ИЮНЬ 2024

SBER PRIVACY JOURNAL



Алексей
Савичев

Руководитель проекта, главный редактор
Исполнительный директор по направлению «Организация обработки и защиты персональных данных», Центр DPO, Сбер

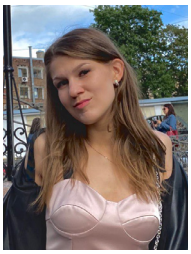
Уважаемые читатели!

Sber Privacy Journal – это корпоративное издание от экспертов Института DPO Сбера о развитии приватности, влиянии новых технологий на персональные данные и о том, как они защищаются в эпоху киберугроз. Наши публикации ориентированы на профессиональное сообщество в области приватности, а также на всех заинтересованных проблематикой обеспечения защиты личных данных.

Для того, чтобы сделать прочтение журнала более приятным и увлекательным, мы исследуем исторические и технологические вопросы, периодически привлекаем внешних экспертов и ученых для подготовки статей, а также предлагаем развлекательный контент в конце каждого выпуска.

Новый, девятый выпуск, Sber Privacy Journal посвящен вопросам эффективного управления данными в эпоху BigData. Приятного чтения!

Редколлегия

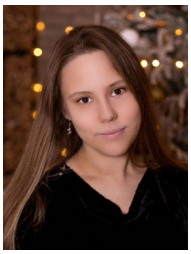


Алёна
Гарцева

Ведущий редактор
Руководитель направления, Центр DPO, Сбер

Евгений
Сердечнюк

Выпускающий редактор
Исполнительный директор, Центр DPO, Сбер



Полина
Сурьянинова

Редактор
Аналитик, Центр DPO, Сбер

Дайджест

Олеся
Сидоренко

Ответственный за дайджест
Эксперт, Центр DPO, Сбер

Дизайн

Ольга
Середа

Дизайнер
Центр DPO, Сбер

СОДЕРЖАНИЕ

Поколения систем управления данными

Взгляд в историю

Андрей Никифоров

Эксперт в области персональных данных, команда DPO Блока B2C, Сбер

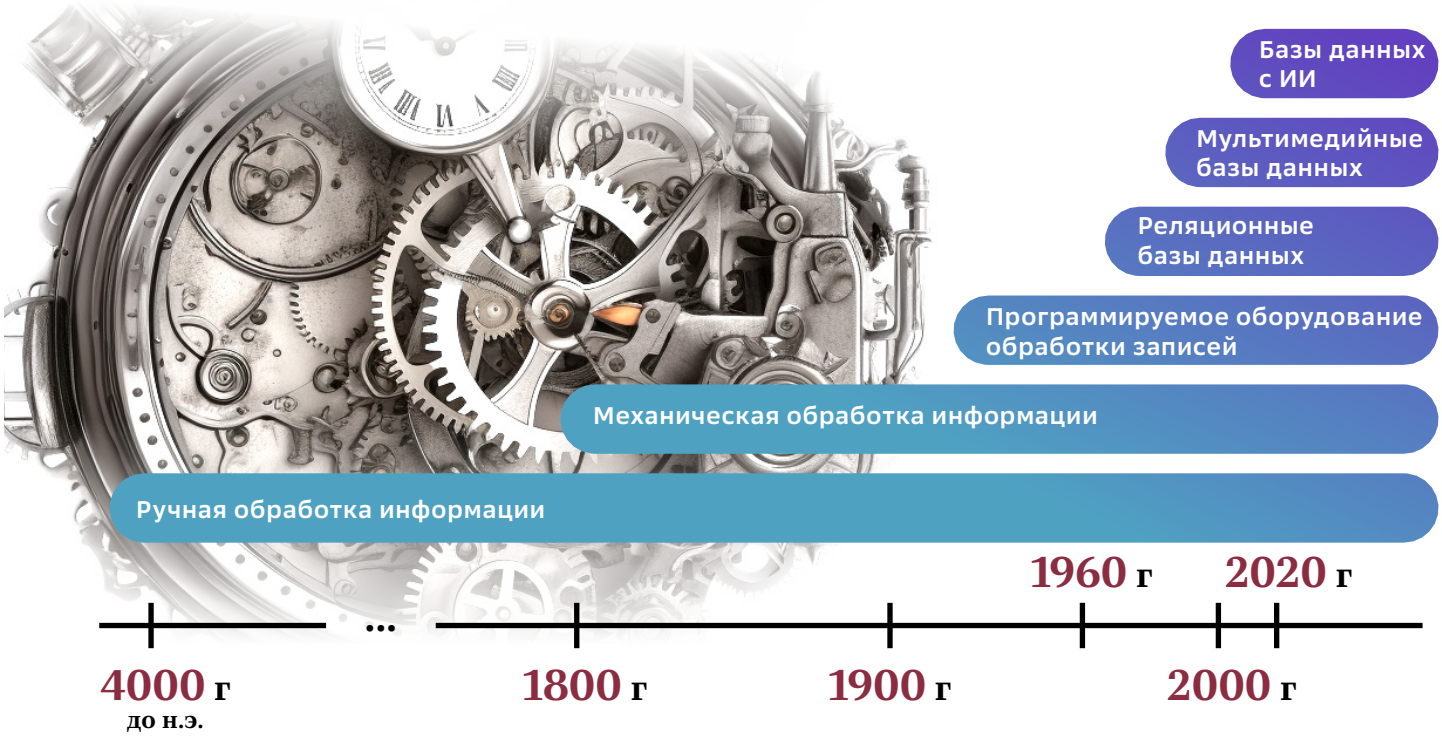
«Всякая перемена прокладывает путь другим переменам».

Никколо Макиавелли[1].

Введение и обзор

История баз данных – это увлекательное путешествие во времени, которое начинается с древних времен и продолжается до наших дней. В этой статье мы рассмотрим основные этапы развития технологий записи данных, начиная с периода до нашей эры и заканчивая современными методами обработки информации.

С самого начала человеческой цивилизации люди стремились сохранять и обрабатывать информацию, в том числе и данные о человеке (на современный лад – персональные данные). С развитием технологий менялись формы, методы и парадигмы обработки данных, однако неизменным было одно – необходимость в более быстрой и масштабной работе с данными, что неудивительно: развитие общества порождает новые человеческие потребности, в том числе в быстром обслуживании, оперативном получении необходимой информации, возможности актуализировать устаревшие данные, включая персональные, и не только.



Изображение 1

Исторические формации систем управления базами данных, подготовлено автором

[1] Макиавелли Н. Государь. История Флоренции (сборник). «ФТМ». «Эксмо». 1532 г. URL: https://fictionbook.ru/static/trials/08/95/37/08953770.a4.pdf (дата обращения: 08.04.2024).

Рассматривая все эти этапы и связанные с ними революционные изобретения, мы захватим в том числе причины перехода на новые ступени – триггеры. Базы данных меняются прямо сейчас и анализ формаций поможет понять происходящее и подготовиться к будущему.

Именно эти технологии позволяют в данный момент управлять нашими персональными данными, понимать, как они защищаются и какие опасности им угрожают. Статья – это приключение через наше прошлое, настоящее и будущее, давайте же отправимся вместе!

Начало: нулевое поколение, ручная обработка информации (4000 г. до н.э. – 1790 г.)

За много лет до того, как базы данных и системы управления ими превратятся в привычные нам сущности и будут, конечно, использоваться для хранения больших массивов данных, в том числе персональных, был качественно иной этап обработки информации – ручной. Он, в свою очередь, имеет два ключевых периода, внутри которых есть свои важные открытия и изобретения, способствовавшие более качественной и эффективной работе с данными.

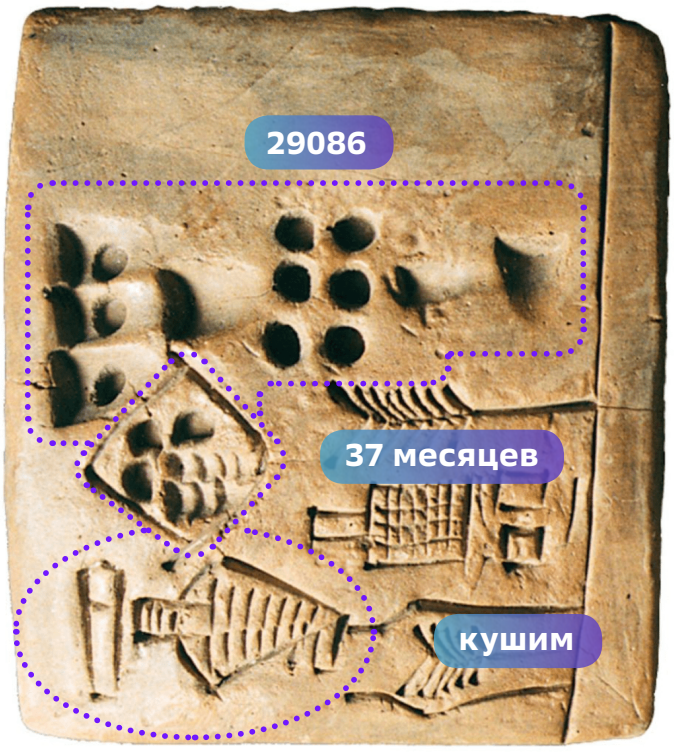
Первым периодом можно считать **устную форму** передачи информации. Таким способом передавались традиции и обычаи, важные события, сведения о социально-значимых, в том числе исторических, личностях: их достижениях и жизни, чертах внешности и характера, привычках и особенностях. Уже тогда люди стремились обработать данные, и особенно какую-то личную информацию, характеризующую предков, с целью передать их потомкам, заложить в основу культуры и социальных норм.

Стоит вспомнить «Илиаду», автором которой считают древнегреческого писателя Гомера. Уже античные ученые считают, что Гомер не пользовался письмом, и произведения его сохранились лишь устно, в памяти певцов, в виде отдельных песен. Несмотря на то, что это одна из множества позиций ученых в «Гомеровском вопросе», устная передача 15700 строк возможна, хотя и потенциально, для современного человека, проблематична[2].

«Илиада» – действительно эпохальное произведение, памятник античного искусства. Но, к сожалению, устная форма фиксации информации и передачи ее из поколения в поколение не позволяла в достаточной мере сохранять повторяемость, точность и поддерживать объем. Более того, устная форма подвержена неточностям и искажениям. Именно поэтому эволюционно человечество пришло к новой, более совершенной форме обработки информации – письменности.

Появление **письменности** можно считать формой передачи информации **второго периода**. Письменность – это метод хранения информации с помощью материальных знаков. Шумерская система письма делала это путем объединения двух типов знаков, которые были выдавлены на глиняных табличках. Один тип знаков представлял числа. Существовали знаки для обозначения 1, 10, 60, 600, 3600 и 36 000. Их шестидесятеричная система передала нам несколько важных наследий, таких как разделение дня на двадцать четыре часа и круга на 360 градусов. Знаки другого типа представляли людей, животных, товары, территории, даты и так далее[3].

Комбинируя оба типа знаков, шумеры сохранили гораздо больше данных, чем запомнил бы любой человек. Обращаясь к первым записям, дошедшим до нас от далеких предков 5000 лет назад, мы видим, так называемые, технические записи. Эти послания гласят: «29 086 мер ячменя, 37 месяцев, Кушим».



Изображение 2

Глиняная таблица с первым зафиксированным именем человека[4]

[2] «Илиада». Большая российская энциклопедия. URL: https://bigenc.ru/c/iliada-da8d26 (дата обращения: 08.04.2024).

[3] Под ред. Юшкевича А. П. История математики с древнейших времён до начала нового времени. М: Наука, 1970. С. 36.

[4] Beer production at the inanna temple in uruk, The Schøyen Collection. URL: https://www.schoyencollection.com/24-smaller-collections/wine-beer/ms-1717-beer-inanna-uruk (дата обращения: 04.03.2024).

Что же кроется за этой фразой? По трактовке историков, это квитанция на несколько партий ячменя. Вероятнее всего, правильно читать это следующим образом: «В общей сложности за 37 месяцев было получено 29 086 мер ячменя. Подпись: Кушим». По этой версии, «Кушим» – первое имя, дошедшее до нас с незапамятных времен[5].

Первые исторические тексты не содержат философских прозрений, поэзии, легенд. Это привычные нам экономические документы, включающие сведения об уплате налогов, накоплении долгов, которые закрепляют их за конкретными людьми.

По своей сути, уже на этом этапе мы видим важность записи и учета информации, которая на сегодняшний день признается в том числе персональными данными, и здесь же можно увидеть **предвестников баз данных**. Собранные в ящики глиняные таблички размещались в специальных помещениях и каталогизировались на отдельных дощечках.

За этим последовало **изобретение более компактных носителей** информации – листов папируса. Научившись обрабатывать пальмовые листы, люди Древнего Египта построили сильную бюрократическую систему учета информации, во главе которой стоял Фараон[6].

Писцы управляли всем этим и составляли мощную иерархию власти. Египтяне проводили учет земли и ее использования. На современный лад это значит, что Египтяне первыми реализовали кадастровый учет. Не менее важными были записи о налогах: как и шумеры, они проводили записи о сборе урожая.

Ярчайший пример так называемой базы данных, содержавшей огромное, даже по сегодняшним меркам, количество личной информации, это сведения о переписи населения. В Циньских картах, захваченных основателем династии Хань Лю Баном, содержались данные в том числе о численности населения. Это дает основание полагать, что уже тогда проводились переписи населения империи, а после воцарения Хань они стали регулярными.

В первой задокументированной демографической сводке, которая относится ко 2 г. н.э., содержатся данные о численности населения отдельных округов и наследственных владений[7].

Появление письменности привнесло количественный скачок в создании баз данных: их прародители стали появляться по всему миру. Целью создания таких объектов хранения информации были систематизация информации, включая личные данные владельцев земель, жителей определенных территорий, и возможность управления ею – настолько, насколько это было возможно в то время. До баз данных в современном понимании далеко, но первые шаги к ним совершались уже тогда.

Качественный скачок произошел в период **изобретения печатного станка**. Он максимально приближает нас к первой формации, хотя все же относится к ручной форме обработки данных.

Массовость в печать привнес в XV веке немецкий ювелир Иоганн Гуттенберг. Он впервые применил технологию печати в Европе. Отлив свинцовые трафареты букв, он получил возможность ставить их отпечатки на бумагу. Всего он отлил не менее пяти разных шрифтов и заложил основу книгопечатания. Благодаря Гуттенбергу у книг появилось понятие экземпляра, а Библия Гуттенберга, первая печатная книга в истории, была затем отпечатана 180 раз[8].

Таким образом, весь нулевой этап, содержащий описанные подэтапы, свидетельствует о важной закономерности – с развитием человечеству необходимо было придумывать новые способы обработки информации, в составе которой были личные данные людей тех времен, с целью повысить качество записи, репликации, хранения, передачи информации, наконец, обеспечить в каком-то смысле управление ею. Неудивительно, что с приходом автоматизации эта закономерность не только не потеряет актуальность, а даже заставит системы управления базами данных развиваться экспоненциально.

[5] Harari Y. N. Sapiens: a brief history of humankind. UK: McClelland & Stewart, 2014. С. 97. URL: https://pdflake.com/wp-content/uploads/2022/01/Sapiens-Book-PDF.pdf (дата обращения: 08.04.2024).

[6] Малыхина И. В. О развитии государственного учёта в Древнем Египте. Статистика и экономика. 2015. №4. URL: https://cyberleninka.ru/article/n/o-razvitii-gosudarstvennogo-uchyota-v-drevnem-egipte (дата обращения: 08.04.2024).

[7] Дикарев А.Д. Население Китая в эпоху Хань (206 г. до н. э. - 220 г. н.э.). Общество и государство в Китае. 2011. № 1. URL: https://cyberleninka.ru/article/n/naselenie-kitaya-v-epohu-han-206-g-do-n-e-220-g-n-e (дата обращения: 08.04.2024).

[8] Библия Гутенберга, Российская государственная библиотека. URL: https://www.rsl.ru/ru/about/projects/biblia-gutenberga-2019/exhibits/bibliya-gutenberga (дата обращения: 08.04.2024).

Первое поколение: начало автоматизированной обработки (1790-1950 гг.)

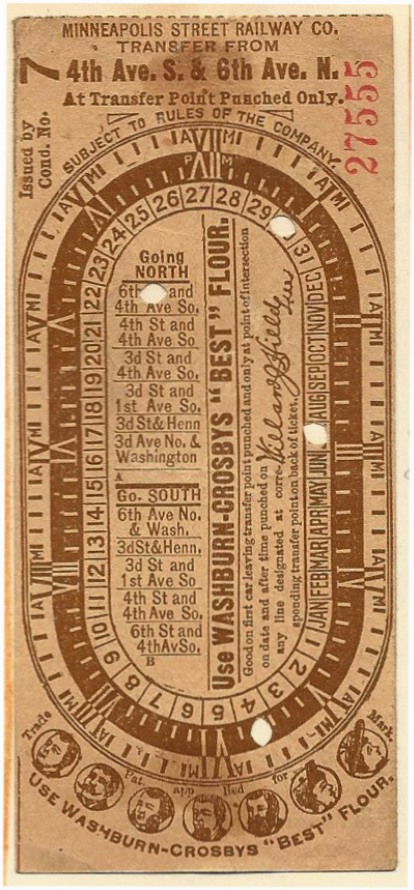
Впервые автоматизированная обработка данных появилась приблизительно в 1790 году, когда Жозефу Жаккарду впервые пришла идея автоматизированного производства ткани при помощи перфокарт. На ткацком станке использовались **сменные перфокарты**, которые управляли плетением ткани таким образом, что любой желаемый узор можно было получить автоматически[9]. Позже, в 1832 году, Семеном Николаевичем Корсаковым была представлена аналогичная технология, использовавшаяся уже для создания машин, способных обрабатывать информацию: решать задачи поиска, сравнения и классификации[10].

По мнению Корсакова, специальные машины могли бы усилить возможности разума. Именно поэтому он предлагал записывать и затем обрабатывать информацию механически, то есть независимо от человека. Они бы могли стать помощниками человека в профильных задачах, а перфокарты тогда стали бы в своем роде базами знаний, в которых записывались лекарства и соответствовавшие им случаи применения. Однако машины не были применены на практике, поскольку комиссия Петербургской академии наук отклонила данное изобретение – хотя, кто знает, возможно, если бы не было этого отказа, перфокарты уже тогда начали бы использовать не только для фиксации случаев применения лекарственных препаратов, но и диагнозов пациентов. Даже возможно, события романа Н.В. Гоголя «Мертвые души», происходившие как раз в тот период времени, развернулись бы совершенно иначе, ведь Чичикову не удалось бы так искусно подтасовывать крестьян!

Однако развитие автоматизированной обработки информации не стояло на месте. В конце концов, несмотря на длительное затишье, перфокарты стали основным средством фиксации информации того времени. Так, в 1890 году электромеханические табуляторы Германа Холлерита, основателя компании IBM, преуспели в масштабной обработке информации, включая персональные данные граждан, и помогли провести перепись населения в США[11]. Благодаря прогрессивным для того времени технологиям данные были обработаны всего за один год, тогда как предыдущая перепись обрабатывалась восемь лет.

Интересный факт

Наблюдая за тем, как кондуктор железной дороги считывает заранее пробитые отверстия в билете, Герман задумал реализовать подобный принцип в большем масштабе. Дело в том, что на железных дорогах того времени пассажиры часто перепродавали свои билеты. Для контроля личности покупателей билетов были введены так называемые «перфорированные фотографии»[12]. На них работник станции пробивал отверстия, которые кратко давали представление о внешности владельца билета, например, форме бороды, цвете глаз, волос, поле и примерном возрасте. Кондуктор мог считать соответствующий набор признаков и произвести идентификацию пассажира. Это не по биометрии человека сверять – раз и готово, тут уметь надо!



Изображение 3

Железнодорожный билет с «перфорированным фото»[13]

[9] Jacquard loom, Britannica. URL: https://www.britannica.com/technology/Jacquard-loom (дата обращения: 08.04.2024).

[10] Александр Михайлов об интеллектуальных машинах Корсакова 1832 года, Пиренеи. URL: https://all-andorra.com/ru/korsakov_mashines/ (дата обращения: 08.04.2024).

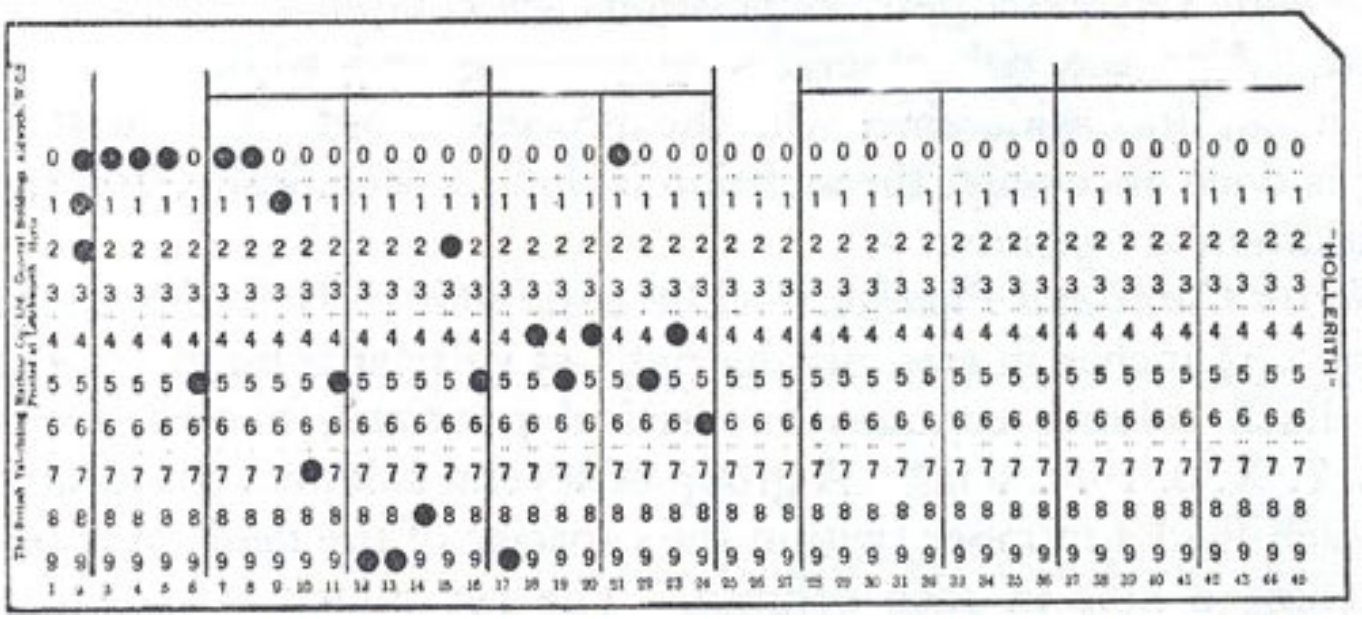
[11] First in the Path of the Firemen, National Archives. URL: https://www.archives.gov/publications/prologue/1996/spring/1890-census (дата обращения: 08.04.2024).

[12] Did you know that punched railroad tickets were the forerunners of computers? Central Pacific Railroad Photographic History Museum. URL: http://cprr.org/Museum/Books/Patton_Made_in_USA.html (дата обращения: 08.04.2024).

[13] «Photographic» punch ticket, Central pacific railroad discussion group. URL: https://discussion.cprr.net/2011/03/question-from-ross-perots-office-in.html (дата обращения: 08.04.2024).

При переписи система Холлерита содержала запись для каждого члена каждой семьи – просто вдумайтесь и оцените, насколько люди в то время стремились к систематизации и управлению имеющимися объемами информации. Запись данных представлялась в виде двоичных структур на перфокарте, которая была размером с банкноту доллара США (Изображение 4).

Записывалась информация об имени, фамилии, поле, возрасте, о расе, умении писать или читать, а также об участии на одной из сторон в Гражданской войне. Машины сводили подсчеты в таблицы по жилым кварталам, территориальным и административным округам и штатам.

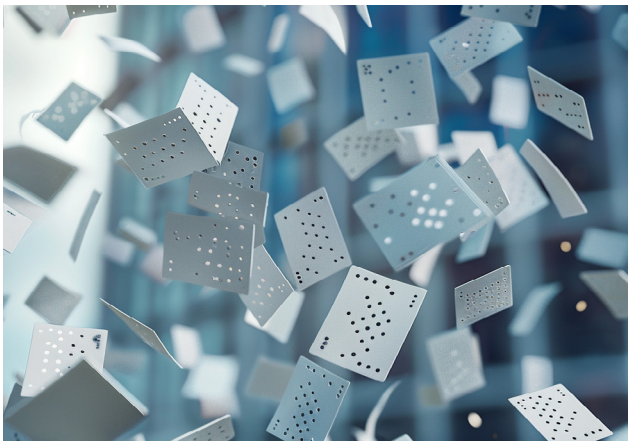


Изображение 4

Перфокарта Холлерита для переписи населения США 1890 года[14]

Холлерит основал компанию по производству оборудования для записи данных на карты, сортировке и составлению таблиц. Бизнес Холлерита в конце концов привел к возникновению International Business Machines. Эта небольшая компания IBM процветала в период от 1915 до 1960 года как поставщик оборудования регистрации данных для бизнеса и правительственных организаций.

К сожалению, первоначальные данные переписи 1890 года больше недоступны. Почти все таблицы численности населения были повреждены в 1921 году во время пожара в здании Министерства торговли в Вашингтоне. Увы, базы данных того времени оказались не стойкими к таким негативным внешним факторам. В свою очередь, сохранившиеся в пожаре данные были уничтожены Правительством. Можно даже сказать, что это один из первых громких случаев уничтожения информации, среди которой было множество персональных данных. В декабре 1932 года, следуя стандартным процедурам ведения учета, главный клерк Бюро переписи населения направил библиотекарю Конгресса список документов с истекшим сроком хранения и подлежащих уничтожению, включая первоначальные списки переписи 1890 года. Бюро попросило библиотекаря указать какие-либо записи, которые следует сохранить для исторических целей, но он не сделал этого. Конгресс санкционировал уничтожение этого списка записей, а сохранившиеся первоначальные записи были уничтожены по распоряжению Правительства к 1934 году[15].



[14] The Hollerith system stored data in the form of round holes in a 45-column card, Early office museum. URL: https://web.archive.org/web/20210506044057/http://www.officemuseum.com/data_processing_machines.htm (дата обращения: 08.04.2024).

[15] 1890 United States Census, Fandom. URL: https://familypedia.fandom.com/wiki/1890_United_States_Census (дата обращения: 08.04.2024).

Изображение 5

Около 4 Гб данных в виде перфокарт IBM в Федеральном центре записей в Александрии, штат Вирджиния, ноябрь 1959 года[16]

К 1955 году у многих компаний имелись целые этажи, предназначенные для хранения перфокарт и управления информацией, которая на них фиксировалась, во многом подобно тому, как в шумерских архивах хранились глиняные таблицы. На других этажах размещались шеренги перфораторов, сортировщиков и табуляторов. Модифицированную версию перфоленты продолжали использовать до восьмидесятых годов прошлого века. Их активно применяли, например, на заводах для станков с числовым программным управлением.

Уже тогда большие компании обрабатывали и производили миллионы записей каждую ночь. Это было бы невозможно при использовании ручных методов обработки. Тем не менее было ясно, что наступает время, когда новая технология вытеснит перфокарты и электромеханические компьютеры.

Второе поколение: устройства с хранимыми программами (1950-1970 гг.)

IBM не стала останавливаться на достигнутом и в 1943 году создала **первый программируемый компьютер** Марк I[17]. Однако он все еще совмещал механические и электронные компоненты. В ENIAC (Электронный числовой интегратор и вычислитель) уже использовалась технология электровакуумных ламп. Он был создан во время Второй мировой войны по заказу армии США для расчета баллистических таблиц, но собрали его только к зиме 1946 года[18].

Поэтому первенство в гонке ламповых компьютеров тогда сохранял британский Colossus, который имел одну задачу – расшифровать немецкие сообщения[19]. Эта машина была схожа с немецкой Enigma, а к концу войны на вооружении Британии стояло 10 таких машин. После войны все они были уничтожены, а данные о них засекречены. Несмотря на прогресс в скорости обработки информации, машина не могла использоваться в других областях, кроме дешифровки.

Прогресс в области записи информации принесли **магнитные ленты**, популяризация которых произошла благодаря UNIVAC[20], одному из самых ранних коммерческих компьютеров. Подобные ему компьютеры были разработаны под технологию магнитных лент и могли обрабатывать гораздо больше данных при меньшей площади, чем их предшественники.

Их отличительной особенностью были программы, записанные в память самого устройства. Сортировка, анализ и обработка[21] информации начали требовать применения языков программирования, а бизнес-использование привело к созданию предшественников стандартных офисных пакетов приложений. Они помогали с общей бухгалтерией, расчетом заработной платы, ведением инвентаризации, банковской деятельностью – например, с проведением платежей. Более того, программы, используемые тогда, применяются, например, в Америке до сих пор и позволяют обрабатывать 80% транзакций по кредитным картам[22].

[16] Punched cards stored in a U.S. National Archives warehouse. Wikimedia commons. URL: https://commons.wikimedia.org/wiki/File:IBM_card_storage.NARA.jpg (дата обращения: 11.03.24).

[17] Harvard Mark I. Britannica. URL: https://www.britannica.com/technology/Harvard-Mark-I (дата обращения: 08.04.2024).

[18] ENIAC. Britannica. URL: https://www.britannica.com/technology/ENIAC (дата обращения: 08.04.2024).

[19] Colossus. Britannica. URL: https://www.britannica.com/technology/Colossus-computer (дата обращения: 08.04.2024).

[20] UNIVAC. Britannica. URL: https://www.britannica.com/technology/UNIVAC (дата обращения: 08.04.2024).

[21] В этом контексте под обработкой подразумевается процесс изменения или преобразования информации в отличие от трактовки термина Законом № 152-ФЗ.

[22] Importance and Application of COBOL in Banking Sectors, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT). URL: https://ijarsct.co.in/Paper5405.pdf (дата обращения: 08.04.2024).

Изображение 6

Приблизительно 13,84 – 104,5 Гб данных, записанных на магнитные ленты в библиотеке магнитных лент Национального центра океанографических данных[23], расчеты автора[24]

Запись на магнитной ленте содержала уже не «единицу» или «ноль», а целый файл. Программы последовательно читали несколько входных файлов и производили на выходе новые. Системы пакетной обработки транзакций сохраняли, например, операции на картах или лентах и собирали их в пакеты для последующей обработки. Раз в день эти пакеты операций сортировались. Затем они сливались с хранимой на ленте намного большей по размерам базой данных – основным файлом – для производства нового.

Несмотря на большую, даже по сегодняшним меркам, вместимость, магнитные ленты предоставляли последовательный доступ к данным. Это означает, что для доступа к файлу, который находится в самом конце, нужно было прочитать всю ленту. Стоит ли говорить, что этот способ обработки информации был медленным.

Кстати, вопрос для экспертов:

«Как с помощью простого карандаша найти определенный фрагмент данных на кассете?»

Не отвечайте, пусть это останется нашим профессиональным секретом.

Пакетная обработка позволяла очень эффективно использовать компьютеры, но обладала на тот момент двумя серьезными ограничениями. Ошибки в транзакциях не распознавались до следующей обработки основного файла, а на ее исправление могло потребоваться еще несколько дней. Следовательно, бизнес не знал текущего состояния данных – поскольку транзакции реально обрабатывались не в моменте, а лишь в определенное время.

Может показаться удивительным, но технология магнитных лент, несмотря на описанные выше недостатки, актуальна и в наше время. Ее востребованность сейчас во многом связана с тем, что удалось нивелировать описанные выше ограничения: новые картриджи поддерживают файловую систему LTFS (Linear Tape File System), которая индексирует содержимое ленты, что позволяет ускорить чтение данных и создает в некотором роде иллюзии произвольного (свободного) доступа к ним. Что не мало важно, они считаются очень надежными и сравнительно дешевыми. Сроки хранения данных на пленке составляют от 15 до 30 лет, а стоимость в закупке меньше, чем у HDD дисков. С каждым годом прирост уплотнения записи данных на лентах стабильно выше такого же показателя у аналога. Поэтому магнитная лента становится даже более востребованной, чем превосходящие ее в других аспектах аналоги. Хранение важной информации на магнитных лентах становится все шире: научные данные о физике частиц, национальные архивы, искусство и наследие, в конце концов, персональные данные. Google сообщал[25], что записывает резервные копии на пленку, в том числе даже для данных сервиса электронной почты. В облачной платформе компании Microsoft Azure говорят[26], что используют пленочное оборудование для дополнительной защиты своих пользователей от внешних атак и посягательств на данные.

Однако, несмотря на эти плюсы, самое популярное использование накопителей с лентой – создание офлайн-копии данных при резервном копировании баз данных. Попросту говоря, копии хранятся, будучи отключенными от средств вычислительной техники, в отдельном хранилище. Это самая эффективная защита, например, от программ-вымогателей, сбоев, выхода из строя оборудования. Если информацию зашифровали

[23] Dorothy Whitaker works in the National Oceanographic Data Center (NODC) magnetic tape library. Wikimedia commons. URL: https://commons.wikimedia.org/wiki/File:NDOC_magnetic_tape_library.tiff (дата обращения: 11.03.2024).

[24] Произведены путем подсчета видимых на снимке магнитных лент и умножения на средний объем данных, которые вмещались на них в том периоде.

[25] Gmail back soon for everyone. Official Gmail Blog. URL: https://gmail.googleblog.com/2011/02/gmail-back-soon-for-everyone.html (дата обращения: 11.03.2024).

[26] Why is Microsoft Azure choosing tape. Fujifilm. URL: https://datastorage-na.fujifilm.com/why-is-microsoft-azure-choosing-tape/ (дата обращения: 11.03.2024).

или удалили, или она стала недоступна по причине сбоев – «из-за плинтуса» поднимается картридж с бэкапом. И вот уже, без потраченных нервов, не нужно восстанавливать данные с уцелевших после атаки или сбоя носителей или платить аферистам за разблокировку.

Тем не менее, эволюция систем продолжалась и актуальные проблемы все так же требовали решения. Несовершенства были исправлены на следующем шаге развития систем управления базами данных – оперативных системах. Это позволило не только ускорить обработку информации, но и иметь к ней доступ в реальном времени.

Третье поколение: оперативные сетевые базы данных (1965-1980 гг.)

Для множества бизнес-процессов необходимы данные в текущем моменте. Устаревшая информация может привести к серьезным последствиям. Поэтому компании, обслуживающие клиентов, попросту не могут использовать неактуальную информацию – им нужен немедленный доступ к текущим данным.

В 1960-х лидирующие компании из нескольких областей индустрии начали вводить в использование **системы баз данных с оперативными транзакциями**, где они обрабатывались в интерактивном режиме[27].

Оперативная обработка транзакций дополняла возможности пакетной, за которой оставались задачи фонового формирования отчетов. Оперативные базы данных хранились уже на магнитных барабанах, которые обеспечивали доступ к любому элементу данных за доли секунды. Эти устройства и программное обеспечение управления данными давали возможность программам считывать несколько записей, изменять их и затем возвращать новые значения пользователю.

С развитием приложений возникла потребность связывать две или более записей. Работавший в General Electric Чарльзом Бахман, возглавил работу группы Data Base Task Group, результатом чего стал прототип системы навигации по данным. За руководство над группой, разработавшей стандартный язык определения данных и манипулирования данными, Бахман получил Тьюринговскую премию. В лекции, которая приурочена премии, он описал переход от компьютерно-ориентированного подхода к базам данных[28]. Благодаря этому кардинально новому подходу в сообществе родились решения проблем с базами данных, в частности, предоставление различного уровня доступа к записям, и ускорилось освоение новых структур данных. В последующем это новое понимание приведет к новым решениям наших проблем с базами данных.

В оперативных сетевых базах данных поддерживались три вида схем данных:

1. **Логическая**. Определяет глобальный проект записей и связей между ними.
2. **Физическая**. Описывает физическое размещение записей на устройствах и в файлах, а также индексы, нужные для поддержки логических связей.
3. **Подсхема**. Раскрывает только часть логической схемы, которую использует программа.

Все схемы в совокупности позволяли обеспечивать независимость данных для того, чтобы предоставить многопользовательский режим их использования. Даже сейчас многие программы, написанные тогда, все еще работают с использованием той же самой подсхемы, хотя логическая и физическая схемы поменялись. Оперативные системы решили проблему одновременного выполнения многих транзакций над базой данных, совместно используемой многими пользователями.

Четвертое поколение: реляционные базы данных и архитектура клиент-сервер (1980-1995 гг.)

Несмотря на успех сетевой модели данных, многие разработчики программного обеспечения понимали, что интерфейс был тяжелым и трудным в освоении. Кроме того, из-за того, что логика процедуры выборки данных зависела от их физической организации, сетевая модель не была полностью независимой от приложения – соответственно, при использовании такой модели, если необходимо было изменить структуру данных, требовалось изменить и приложение, что, конечно же, было достаточно ресурсозатратно.

В 1970 году Эдгар Кодд написал первую работу по **реляционной модели данных**: «A Relational Model of Data for Large Shared Data Banks»[29]. Затем, в 1981 году, он разработал реляционную модель данных и реляционную алгебру, за что также получил премию Тьюринга. В лекции Кодд описал реляционную модель, идея которой состояла в том, чтобы единообразно представлять и сущности, и связи. Эта модель обладала унифицированным языком для определения, навигации и манипулирования данными. Операции в ней применялись ко множеству показателей целиком и производили свой результат от всех них сразу. Благодаря этому программы для решения задач управления записями становились короче и проще.

[27] Сетевые базы данных. Timetoast. URL: https://www.timetoast.com/timelines/-8a2a1481-f111-4af5-9473-ab26a8137639 (дата обращения: 08.04.2024).

[28] Bachman C. The Programmer as Navigator. Communications of the ACM, V. 16, N. 11, 1973. URL: https://people.csail.mit.edu/tdanford/6830papers/bachman-programmer-as-navigator.pdf (Дата обращения: 08.04.2024).

[29] Любченко Д.П. История возникновения баз данных. Т: Вестник науки №10 (19) том 3, 2019. С. 88.

Помимо повышения продуктивности программистов и простоты использования реляционная модель оказалась хорошо пригодной к использованию в архитектуре клиент-сервер, а также к параллельной обработке и графическим пользовательским интерфейсам.

К 1990 году реляционные системы стали более популярными, чем предшествовавшие им навигационные системы, ориентированные на наборы записей. Между тем, для многих корпораций предки реляционных баз данных все еще были более привычны и актуальны. С годами такие корпорации построили громадные приложения и не могли легко перейти к использованию реляционных систем.

Пятое поколение: мультимедийные базы данных (1995-2020 гг.)

Несмотря на то, что реляционные системы сделали много для развития отрасли, исследовательское сообщество стало рассматривать вопросы, выходящие за рамки существующей модели.

Традиционно существовало четкое разделение программ и данных. Такой подход хорошо работал, пока речь шла только о числах, символах, массивах, списках или множествах записей. С появлением новых приложений это разделение стало проблематичным. Приложениям требовалось соотнести данные с их поведением. Так, если данные представляли сложный объект, то есть содержащий помимо текста изображение или звук, то методы поиска, сравнения и манипулирования ими становились специфичными.

Тот, кто желает использовать иные данные, к примеру, специалист в своей прикладной области, должен определить и разделить их на собственные типы. Географы смогут искать и кодировать карты, копирайтеры – индексировать и проводить выборку текстов, дизайнеры – использовать библиотеки типов для работы с изображениями, сотрудники компании-оператора персональных данных – производить идентификацию субъекта персональных данных.

Однако и при этой формации есть проблема: для разработки описанных типов требуется хорошая модель и унификация процедур и данных. Быстрое распространение интернета особенно обострило этот вопрос. Не углубляясь в детали его решения, стоит подчеркнуть, что эти базы данных призваны хранить больше, чем только числа и текстовые строки. Они используются для хранения большого количества объектов, которые мы видим в интернете, и связей между ними.

Унификация процедур и данных расширяет традиционную вычислительную модель клиент-сервер в двух интересных направлениях:

1. Активные базы данных

2. Поток операций

Активные базы данных самостоятельно выполняют задачи при изменении в них. Другими словами, холодильник сам будет заказывать молоко, когда оно закончится. Холодильник в этом примере – это так называемая база данных, молоко – показатель, который меняется в процессе. Используя этот принцип возможно реализовать различные сценарии, которые переносят логику из приложений к данным, а также делают базы данных самоуправляемыми.

Поток операций – это последовательность задач, которые должны быть выполнены. Например, обычный заказ на маркетплейсе – это множество шагов от выбора товара, заполнения личных данных (имени получателя, адреса доставки, номера телефона и не только) до его оплаты. Такие системы не менее распространены для различных сценариев и обработки данных.

Потребности в увеличении скорости и масштабов обработки приводят к созданию колоссальных проектов в области управления данными. Система Earth Observation System Data and Information System (EOSDIS)[30] разрабатывается NASA и предоставляет комплексные возможности для управления данными о Земле из различных источников – спутников, самолетов, полевых измерений и т.д. Объем базы данных на 1 апреля 2024 года составлял 102,5 петабайт, а растет она в среднем на 91,64 терабайт в день[31]!

Мультимедийные базы данных представили новые решения для обработки, в том числе хранения, информации, что позволило эффективно управлять большими объемами данных различных типов. Переход к этой формации был все также обусловлен постоянной эволюцией систем управления базами данных и необходимостью удовлетворения потребностей современных пользователей. Они позволяют не только хранить и обрабатывать данные, но и предоставлять доступ к ним в удобной и интерактивной форме. Являясь ключевым элементом в создании современных информационных систем, они обеспечивают возможность эффективного использования данных в различных областях деятельности.

[30] Earth Observing System Data and Information System (EOSDIS). EARTHDATA. URL: https://www.earthdata.nasa.gov/eosdis (дата обращения: 08.04.2024).

[31] System Performance and Metrics. EARTHDATA. URL: https://www.earthdata.nasa.gov/eosdis/system-performance-and-metrics (дата обращения: 08.04.2024).

Шестое поколение: базы данных и искусственный интеллект (с 2020 г.)

Революция в области **искусственного интеллекта** меняет все отрасли нашей жизни, обещая невероятные инновации с одной стороны и сталкивая нас с новыми вызовами с другой. Не исключением стала и сфера обработки данных, включая, конечно же, персональные данные, в которой появляются новые вызовы в аспекте доступности данных и их безопасности. Постоянное увеличение количества данных ставит эффективную обработку в приоритет для программ, на основе больших языковых моделей, генеративного ИИ и семантического поиска.

В основе новых технологий лежат уже векторные представления данных – эмбеддинги (англ. embeddings). Ими достаточно сложно управлять, так как они охватывают множество атрибутов или характеристик. В области искусственного интеллекта и машинного обучения эти характеристики представляют различные измерения данных, необходимые для обнаружения закономерностей, взаимосвязей и базовых структур.

Для удовлетворения уникальных требований к обработке этих вложений необходима специализированная база данных. **Векторные базы данных** специально созданы для обеспечения оптимизированного хранения и запросов векторов, сокращая разрыв между традиционными базами данных и самостоятельными векторными индексами, а также предоставляя ИИ-системам инструменты, необходимые для успешной работы в этой среде, нагруженной данными.

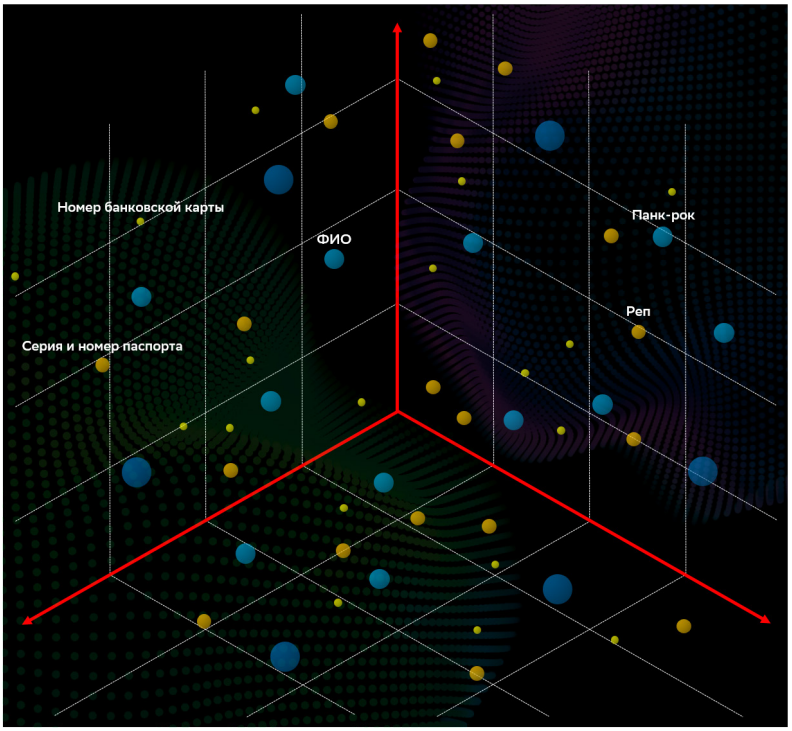
Векторная база данных – это специализированный тип базы данных, который хранит данные в виде многомерных векторов, каждый из которых представляет определенные характеристики или качества. Эти векторы могут иметь различное количество измерений в зависимости от сложности данных.

Основное преимущество векторной базы данных заключается в ее способности эффективно и точно извлекать данные на основе близости или сходства векторов. Это позволяет выполнять поиски на основе семантической и контекстуальной значимости и не полагаться только на точные совпадения или заранее определенные критерии, как в традиционных базах данных.

Важно рассмотреть, как именно векторное представление данных, эмбеддинги произвели революцию. Это массивы чисел, которые содержат семантическое значение объектов данных. Представление сложных данных, таких как слова или текст, в значимые числовые последовательности является амбициозной задачей.

Эмбеддинги кодируют семантическое значение объектов относительно друг друга. Похожие объекты группируются близко в векторном пространстве, что означает – чем ближе два объекта, тем больше они похожи. Например, рассмотрим словосочетания «номер банковской карты» и «серия и номер паспорта». Они находятся близко друг к другу, потому что зачастую к определенному паспорту присоединяются определенные номера банковских карт. «ФИО» также имеет связь, так как эти данные могут быть связаны и с номером банковской карты, и с серией и номером паспорта.

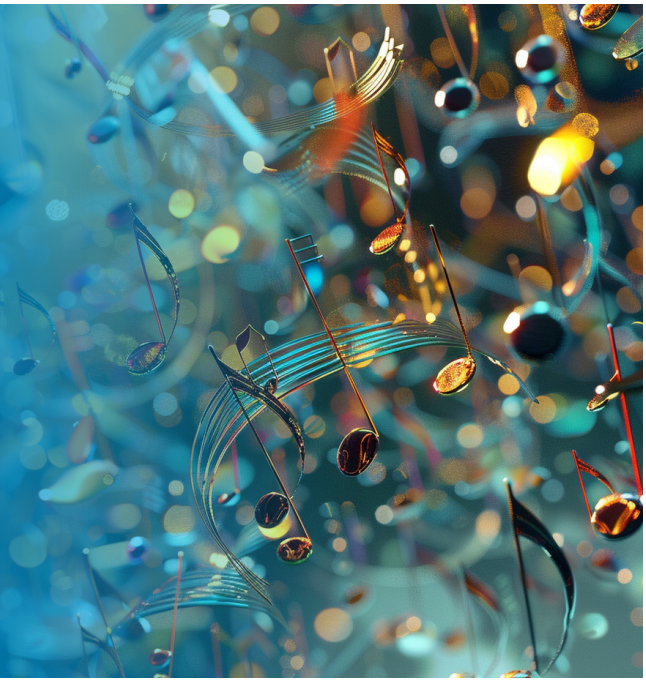
С другой стороны, слова, представляющие жанры музыки, такие как «панк-рок» и «реп», находятся дальше от терминов, относящихся к персональным данным, формируя отдельный кластер в векторном пространстве. Тем не менее их также можно успешно использовать для персонализации приложений.



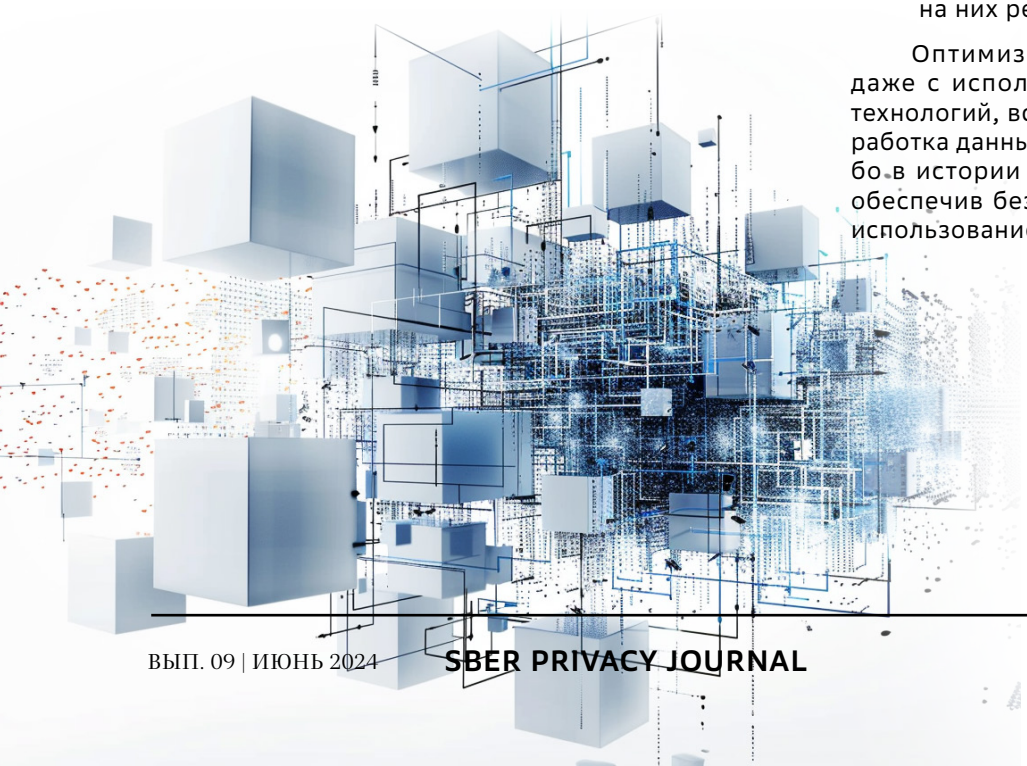
Изображение 7

Векторное представление данных, подготовлено автором

Яркий пример успешной реализации такого использования – стриминговые сервисы, например, музыкальные. Используя векторное представление данных от различных жанров музыки до конкретных произведений, можно получить определенное числовое значение. Затем, предварительно проанализировав поведение пользователя: его регистрационные данные, пройденные тесты на предпочтения в музыке, «лайкнутые» или «дизлайкнутые» треки, – можно составить второй вектор. Результатом сравнения двух этих величин будет прогноз: понравится ли пользователю композиция или нет. Таким образом, сервис может постоянно обновлять список треков для ознакомления, подбирать давно не прослушиваемые, знакомить с новыми жанрами с учетом предпочтений пользователя и так далее.



Векторные базы данных играют ключевую роль в индексации векторов, созданных с помощью эмбеддингов. Они позволяют осуществлять поиск похожих ресурсов с помощью соседних векторов. Разработчики используют эти базы данных для создания уникального пользовательского опыта, включая поиск изображений на основе снимков, сделанных пользователями. Автоматизация извлечения метаданных из контента, вместе с гибридным поисковым запросом на основе ключевых слов и векторов, дополнительно повышает возможности поиска. Еще векторные базы данных служат внешними базами знаний для генеративных моделей искусственного интеллекта, так они обеспечивают стабильное взаимодействие с пользователем.



Векторные базы данных оказывают значительное воздействие на различные отрасли:

- **В розничной торговле** они персонализируют рекомендации клиентам, выдавая предложения на основе предпочтений и выбранных им продуктов.
- **В финансовой сфере** на их основе проводится анализ закономерностей рынка и выводятся уникальные стратегии, учитывающие характер инвестора и объективные закономерности рынка.
- **В сфере безопасности** и противодействия мошенничеству векторные базы данных отлично подходят для обнаружения аномалий и выбросов, что позволяет быстрее и точнее на них реагировать.

Оптимизация производительности СУБД, даже с использованием таких революционных технологий, все еще сложная задача, и сейчас обработка данных ускоряется больше, чем когда-либо в истории человечества. Справиться с этим, обеспечив безопасность, приватность и этичное использование данных, наша с вами задача.

Заключение

Вся история развития систем управления базами данных была пронизана главной тенденцией – экспоненциальным переходом от менее эффективных к другим более быстрым и удобным. Рост потребностей в части управления данными каждый раз приводил к появлению новых поколений систем управления базами данных.

В период ручной обработки, несмотря на естественность процесса, нельзя было собирать и преобразовывать большие объемы данных, а сама информация часто могла быть недостоверной.

В следующий, механический, период данные стали записываться гораздо более точно, но еще не так эффективно с точки зрения скорости обработки.

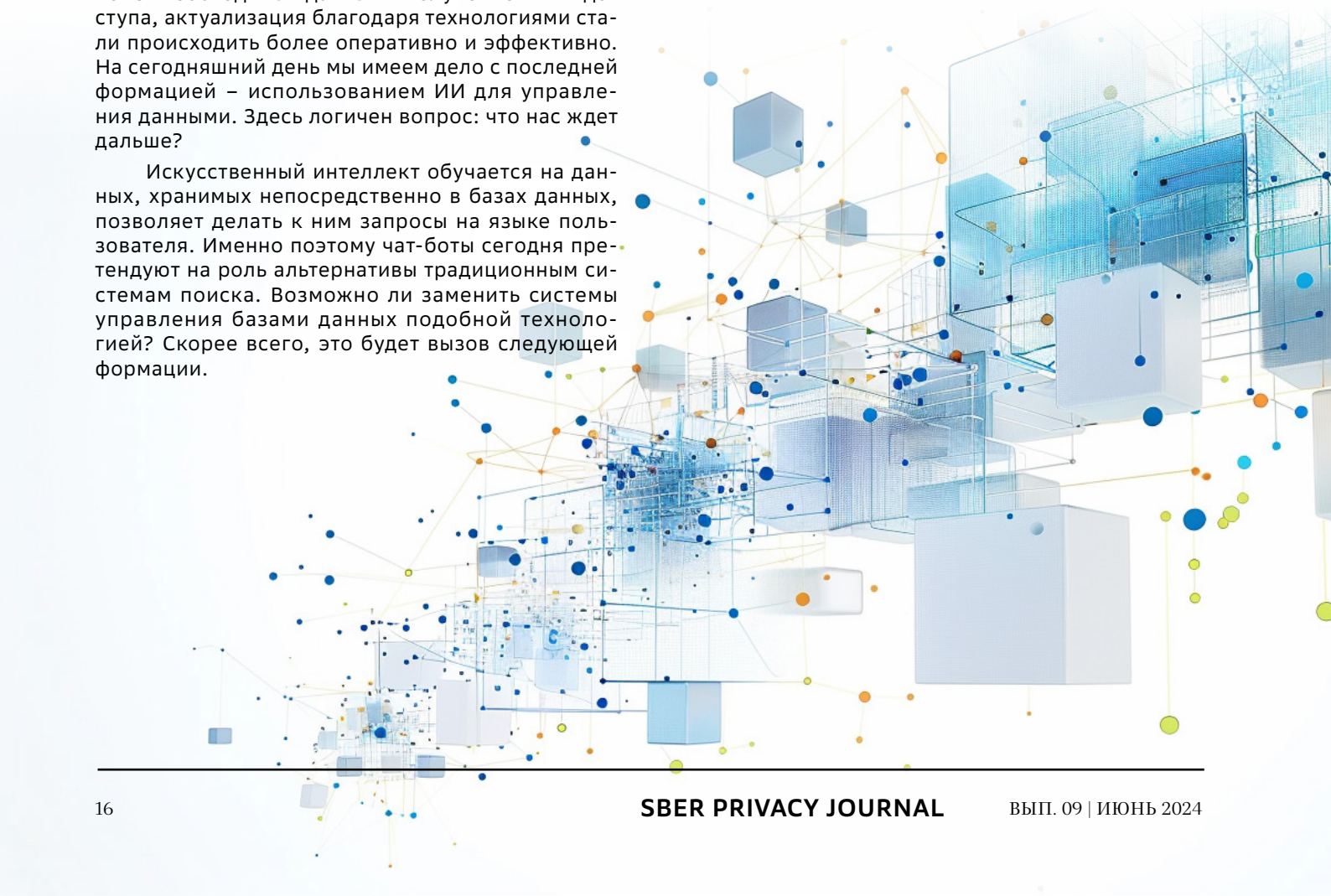
Другие периоды один за одним позволяли увеличить как скорость обработки данных, так и удобство работы с ними. Достаточно вспомнить изменение скорости и качества обслуживания клиентов на каждом этапе, чтобы оценить колоссальную разницу и прогресс. Если раньше процесс изменения клиентских данных предполагал обязательное посещение офиса компании, предоставление в офлайн формате большого количества подтверждающих документов, то теперь для совершения многих действий достаточно нескольких кликов в интерфейсе приложения. И все это потому, что управление данными, включая сбор, поиск необходимых данных и получение к ним доступа, актуализация благодаря технологиями стали происходить более оперативно и эффективно. На сегодняшний день мы имеем дело с последней формацией – использованием ИИ для управления данными. Здесь логичен вопрос: что нас ждет дальше?

Искусственный интеллект обучается на данных, хранимых непосредственно в базах данных, позволяет делать к ним запросы на языке пользователя. Именно поэтому чат-боты сегодня претендуют на роль альтернативы традиционным системам поиска. Возможно ли заменить системы управления базами данных подобной технологией? Скорее всего, это будет вызов следующей формации.

Сейчас перед искусственным интеллектом стоит задача решить проблему «Галлюцинаций», когда он выдает выдуманные ответы или меняет формат выдачи по своей прихоти, а также проблему способности работать «в контексте» (то есть выдавать ответ в рамках предметной области). Но если предметная область достаточно узкая, а обучающая выборка по ней была исчерпывающей и свободной от ошибок, то для описанных ранее задач технология могла бы стать заменой традиционному подходу.

В любом случае описанные преимущества, а также ускоряющееся появление новых поколений систем управления базами данных приведет нас к еще более точной и эффективной системе.

И хотя со стороны защиты персональных данных остаются нерешенными многие вопросы относительно этической составляющей обучения и применения моделей искусственного интеллекта на реальных кейсах с персональной информацией, очевидно: эволюция будет продолжаться. Новые технологии требуют все большего количества этого типа данных для увеличения своей эффективности. Нам остается быть бдительными и не забывать об истории, ведь следующие формации всегда решают актуальные проблемы предыдущих.



Что такое управление данными

и как оно стало основой трансформации цифровых процессов

Олег Беляев

Руководитель направления, команда DPO Блока КИБ, Сбер

«Кто владеет информацией, тот владеет миром».
Н. Ротшильд.

Что такое данные и зачем ими нужно управлять?

С давних времен информация стала важным элементом в жизни и развитии общества. В разное время на разных уровнях, будь то жизнь одной семьи, компании или государства, своевременно полученная и точная информация, соответствующая потребности, помогала достижению высоких результатов, власти, богатства. В истории существует много примеров, когда благодаря информации компании получали конкурентное преимущество, государства занимали определенные позиции на международной арене, но также есть и примеры, когда ложная или неточная информация становилась яблоком раздора и приводила к негативным последствиям. В современном мире важность информации стала еще больше. Как сказал британский математик и эксперт в области анализа данных Клайв Хамби в 2006 году: «Данные – это новая нефть». Эта фраза, ставшая афоризмом, подчеркивает не только саму важность и большую ценность данных, но и то, что важно не только иметь данные, но и уметь их обрабатывать и извлекать выгоду. Характерной чертой 21-го века является **ежегодный рост объемов информации**, обрабатываемой в мире. В докладе «Эпоха данных – 2025» (The Data Age 2025[1]), подготовленном аналитиками компании IDC при поддержке производителя жестких дисков Seagate, авторами дается оценка, что к 2025 году общемировой объем данных вырастет в 10 раз и достигнет 163 зеттабайт (Зб)[2], причем большую часть этих данных будут генерировать предприятия, а не потребители. Такой рост объема данных делает еще более сложным задачу поддержания их точности, актуальности, а также выдвигает более строгие требования к обрабатывающим их системам с целью обеспечения потребности в получении того объема данных, который нужен для конкретной цели их обработки и в тот момент времени, когда эти данные необходимы. Становятся более сложными для решения вопросы: где (в каких системах) хранятся данные, кто эти данные использует, в каких целях, как обеспечивать их конфиденциальность и целостность, своевременную актуализацию, – что приводит нас к необходимости выработки системного, поэтапного подхода, целью которого является структурирование деятельности по работе с данными для достижения задач, указанных выше, а также для обеспечения возможности получать больше выгоды от данных и снижения рисков их некорректной[3] обработки. Таким подходом является **система[4] управления данными**. В data-driven компаниях, основывающих свои бизнес-процессы на данных, обеспечение защиты и выполнение требований законодательства к обработке персональных данных является одной из ключевых задач в выстраивании бизнес-процессов, для достижения которой необходимо так или иначе внедрять практики управления данными. Что они из себя представляют? Далее мы более подробно поговорим о ключевых принципах построения процессов управления данными и их необходимости при развитии и цифровой трансформации современных компаний.

[1] Источник: Data Age 2025. URL: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/workforce/ey-seagate-wp-data-age-2025-march-2017.pdf (дата обращения: 22.04.2024).

[2] В одном зеттабайте содержится 10 в 21-ой степени байтов.

[3] В частности, без наличия законных оснований, в избыточном объеме для заявленной цели, с незаконной передачей третьим лицам и другими способами, нарушающими требования кибербезопасности и применимого законодательства.

[4] Под системой понимается комплекс мероприятий, правил, документов и процессов.

Что такое управление данными?

Рассмотрим ситуацию: индивидуальный предприниматель (далее – ИП) пришел в условный банк, чтобы изменить информацию о документе, удостоверяющем личность (паспорте гражданина РФ, который был заменен на новый по достижении определенного возраста). Статус индивидуального предпринимателя: физическое лицо, осуществляющее предпринимательскую деятельность без образования юридического лица (оговорка об осуществлении предпринимательской деятельности имеет важное значение). Данные в банке были успешно зафиксированы специалистами отдела, который сопровождает продукты корпоративного бизнеса[5]. Однако через несколько дней уже из розничного отдела, который работает с физическими лицами (не корпоративными клиентами), его просят приехать в тот же офис, чтобы сделать скан-копию паспорта, как клиента физического лица (далее – ФЛ). Интуитивно понятно, что такой клиентский путь не способствует поддержанию лояльности клиентов. Почему такое могло произойти?

Как правило, ЮЛ (а также ИП, самозанятых, лиц, занимающихся частной практикой) и ФЛ (без указанных статусов) обслуживают разные подразделения и, соответственно, разные отделы банка и клиентские менеджеры. В нашем примере:

- клиент изначально пришел в банк как ИП к клиентскому менеджеру, который работает с ИП, а не с ФЛ без статуса ИП;
- далее клиент-ИП передал данные о «новом» паспорте этому клиентскому менеджеру, а он, в свою очередь, обновил данные в автоматизированных системах подразделения, отвечающего за клиентов ИП (в автоматизированных системах подразделения, отвечающего за клиентов-ФЛ без статуса ИП, данные обновлены не были как раз потому, что клиентов ИП и клиентов ФЛ-не ИП ведут разные подразделения с разными автоматизированными системами);
- после этого произошло следующее: когда в подразделении, которое работает с клиентами-ФЛ (далее – розничное подразделение), появилась информация из МВД[6] о недействительности «старого» паспорта данного клиента-ФЛ (который уже как ИП приходил в банк, но к клиентским менеджерам, отвечающим за ИП), представитель розничного подразделения связался с клиентом-ФЛ и попросил его прийти в офис с «новым» паспортом, хотя, казалось бы, необходимая информация уже есть в банке.

Такая ситуация стала возможной как раз ввиду того, что в отношении одного и того же клиента, но с разными статусами при обслуживании: 1) клиент-ФЛ и 2) клиент-ИП, – в нашем условном банке могут иметься одни и те же сведения, на основании которых устанавливается его личность (при этом устаревшие сведения своевременно не актуализируются во всех автоматизированных системах). Это свидетельствует о недостаточном внимании к вопросам управления данными.

[5] То есть сопровождает юридических лиц (далее – ЮЛ), ИП, самозанятых, лиц, занимающихся частной практикой.

[6] Министерство внутренних дел Российской Федерации.

Итого, в общем виде можно сказать, что управление данными – это процесс, в рамках которого данные рассматриваются как **ценный актив** компании, направленный на обеспечение необходимых свойств данных на всех этапах их жизненного цикла. Одна из **ключевых задач** системы управления данными – создать условия для выстраивания процессов таким образом, чтобы исключить проблемы, связанные, например, с:

- непониманием того, где находятся данные (в каких автоматизированных системах), которое, очевидно, приведет либо к невозможности найти данные вовсе, либо к потребности проверять каждую автоматизированную систему в попытках отыскать нужную информацию (что крайне ресурсозатратно и малоэффективно);
- невозможностью поддерживать данные в актуальном состоянии, обеспечивать их точность (что важно, как минимум, для того, чтобы исключить ошибки в принятии решений на основании недостоверных данных – например, о выдаче кредита или, напротив, отказе в его предоставлении);
- неготовностью выстроить эффективное взаимодействие между подразделениями компании, обслуживающими одно и то же лицо в различных статусах (например, клиент-ФЛ/клиент-ИП);
- неспособностью выполнять требования к обработке данных на каждом этапе их жизненного цикла (в том числе требования специализированного законодательства о персональных данных);
- неготовностью обеспечивать необходимые объемы вычислительных ресурсов для обработки данных;
- невозможностью выполнять требования законодательства к своевременному уничтожению данных.

С ключевыми проблемами, на решение которых направлено управление данными, определились. Теперь давайте более детально разберемся с основными понятиями системы управления данными и поймем, как ее выстраивать и как контролировать выполнение стандартов управления данными.

Большие данные – вызов для управления данными

Первое важное понятие – «Большие данные»[7]. Прежде, чем углубиться в вопросы управления, давайте уточним, чем именно предстоит управлять.

Большие данные – это термин, применяемый к очень большому объему информации, настолько большому, что работа с ним с применением обычных инструментов и методов невозможна. Такие данные используют, например, для подготовки аналитики, статистики, прогнозов и предсказания решений.

Термин впервые был употреблен редактором журнала Nature Клиффорд Линч в спецвыпуске[8] в 2008 году в разговоре о резком росте объемов информации в мире. К большим данным Линч отнес поток данных более 150 Гб в сутки, однако единого критерия до сих пор не существует. Есть шуточный критерий: если данные не открываются в Excel, то это большие данные. Действительно, имея набор данных такого объема, который невозможно обработать вручную с помощью базовых технических программ, вопросы управления данными обретают иной масштаб и требуют более глобальной и централизованной имплементации процессов управления данными в деятельность компании на всех этапах жизненного цикла данных.

По сути, можно говорить о том, что именно большие данные стали неким «катализатором» выстраивания систем управления данными и решения проблем, возникающих в отсутствие управления.

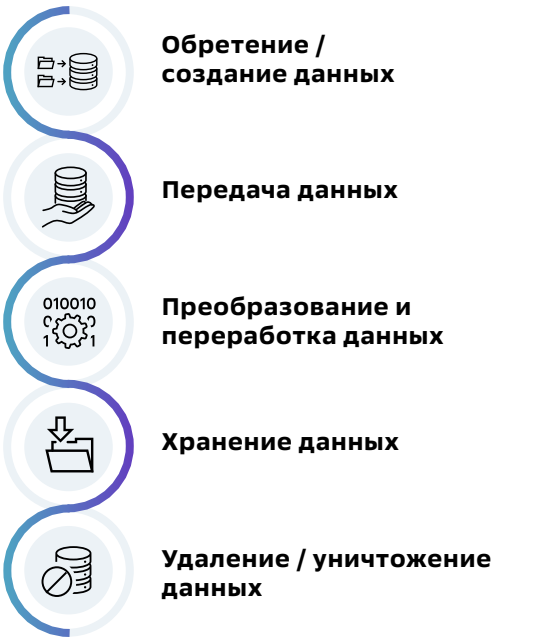
[7] В переводе с английского термина «Big Data».

[8] Nature. Volume 455 Issue 7209, 4 September 2008. URL: https://www.nature.com/nature/volumes/455/issues/7209 (дата обращения: 16.05.2024).

Что такое жизненный цикл данных?

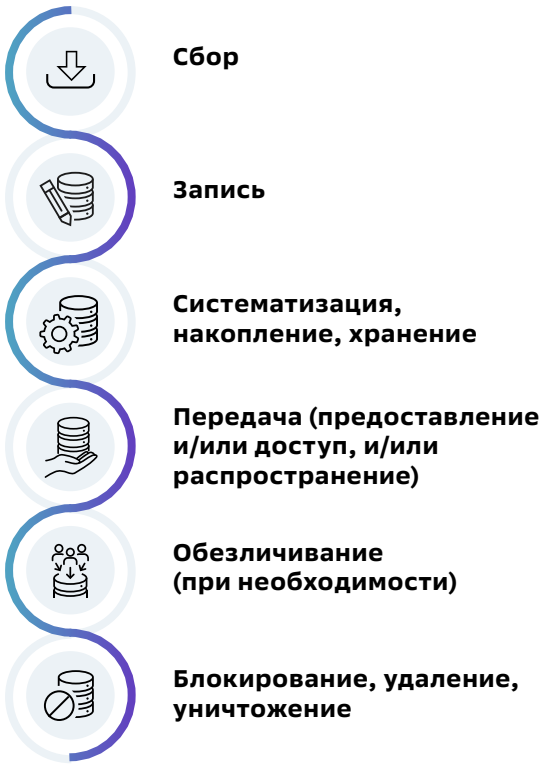
Как у любого объекта управления, у данных есть их **жизненный цикл** – последовательность этапов, стадий, которые данные проходят в процессе своего существования в деятельности компании. В соответствии с методологией DAMA-DMBOK[9], жизненный цикл данных включает этапы, на которых данные создаются или получаются; этапы, на которых осуществляется передача данных; этапы преобразования, обработки и хранения данных, а также этапы уничтожения данных.

Схематично эти этапы можно отобразить следующим образом:



Каждый из приведенных этапов может реализовываться над данными в виде определенных и конкретных действий. Данные могут преобразовываться, объединяться, агрегироваться, дополняться, поэтому фактически внутри каждого этапа существуют свои подэтапы, специфичные для конкретной деятельности, инфраструктуры или вида данных или даже специализированного законодательства.

Например, для сферы персональных данных можно выделить «свой» собственный жизненный цикл, учитывая те действия с данными, которые входят в понятие «обработка персональных данных»:



[9] За основу описываемых подходов в построении системы управления данными взяты понятия и фреймворки одной из самых известных методологий и свода знаний по управлению данными – DAMA-DMBOK. Методология подготовлена Международной ассоциацией управления данными (Data Management Association International, DAMA в 2009 году; DAMA-DMBOK: Свод знаний по управлению данными. Второе издание / Dama International [пер. с англ. Г. Агафонова]. – Москва: Олимп–Бизнес, 2020. – 828 с.

Согласно DAMA-DMBOK, управление данными – это управление их жизненным циклом. На каждом этапе жизненного цикла и для каждого вида данных данные претерпевают то или иное воздействие, которое в конечном счете и подлежит контролю наряду с корректным выстраиванием обработки (то есть каждое действие с данными должно соответствовать применимым требованиям законодательства, кибербезопасности, потребностям компании и др.).

Итого: построение системы управления данными на основе жизненного цикла данных, а точнее для каждого этапа жизненного цикла, **позволяет структурировать** все требования к данным, **разделить их на группы**, применимые к конкретным этапам, и **реализовать** их в определенных бизнес-процессах и автоматизированных системах. Например, процесс сбора персональных данных новых клиентов соответствует этапу жизненного цикла управления данными «Обретение данных» (и этапу «Сбор» для жизненного цикла управления персональными данными). Для этого этапа можно определить ряд требований, таких как: локализация баз данных на территории Российской Федерации (при сборе через интернет), определение состава данных в соответствии с законодательством и последующей целью обработки (с учетом того, какая услуга будет оказана клиенту), определение законных оснований для обработки данных (согласие, договор с субъектом и пр.), определение срока, в течение которого требуется обработка данных. Для этого процесса также можно определить конечный список автоматизированных систем, шаблонов документов и перечень ролей работников, задействованных в процессе. При этом на данном этапе для нас не будут актуальны вопросы передачи данных или их уничтожения.

> Такое гранулярное управление данными является наиболее оптимальным в масштабах крупных компаний при управлении большими данными.

Итак, немного углубившись в суть и объем рассматриваемого вопроса, сформулируем **ключевые аспекты** построения системы управления данными:

1. Необходимость взаимодействия с различными видами данных.
2. Имплементация управления данными в разнообразной инфраструктуре.
3. Работа со множеством бизнес-процессов, использующих данные.
4. Оценка Cost-to-Value и соответствие ожиданиям бенефициаров компании.

Раскроем подробнее каждую тему.

Ключевые моменты в построении системы управления данными

Возможные классификации информации

01 Классификация с учетом правовых режимов информации

С учетом рода деятельности компании внутри ее баз данных, информационных систем, рабочих мест работников и иных информационных ресурсов может обрабатываться информация, которая подразделяется на следующие правовые режимы:

- **в зависимости от категории доступа к ней** – на общедоступную информацию, а также на информацию, доступ к которой ограничен федеральными законами (сюда подпадает информация, отнесенная к государственной тайне, сведения конфиденциального характера, включая персональные данные);
- **в зависимости от порядка предоставления или распространения** – информация, свободно распространяемая; информация, предоставляемая по соглашению лиц, участвующих в соответствующих отношениях; информация, которая, в соответствии с федеральными законами, подлежит предоставлению или распространению; информация, распространение которой в РФ ограничивается или запрещается[10].

Для обработки указанной информации существуют свои правила и требования, которые должны быть реализованы на каждом из этапов жизненного цикла.

Например, при обработке персональных данных необходимо в каждом процессе/направлении деятельности компании, где осуществляется обработка персональных данных:

1. При сборе – как минимум, обеспечить наличие правовых оснований, соответствующих заявленной цели, а также локализацию баз данных в случае сбора данных посредством сети интернет.
2. При записи и систематизации – например, определить форматы места записи данных с учетом необходимости предоставления возможности оперативного обращения к любому требующемуся составу данных.
3. При хранении – определить сроки хранения, места хранения, возможность или невозможность совместного хранения с данными, собранными для обработки в иных (совместимых/несовместимых) целях.
4. При актуализации – определить порядок уточнения персональных данных, в том числе в резервных копиях баз данных.
5. При использовании и передаче – определить необходимый объем предоставляемых данных и наличие правовых оснований на их передачу.
6. При организации доступа к персональным данным – обеспечить предоставление доступа к минимально необходимому объему данных и только тем работникам, которым он требуется для выполнения трудовых обязанностей.
7. В случае прекращения обработки – обеспечить блокирование и уничтожение персональных данных в сроки, установленные 152-ФЗ.

Таким образом, процессы управления данными нужно имплементировать в каждое действие по обработке персональных данных, предусмотренное в 152-ФЗ[11].

[10] Ч. 2, 3 ст. 5 Федерального закона от от 27.07.2006 № 149-ФЗ (ред. от 12.12.2023) «Об информации, информационных технологиях и о защите информации» (здесь и далее по тексту – 149-ФЗ).

[11] В статье приведены некоторые действия по обработке и обращено внимание на одни из ключевых процессов управления в каждом из таких действий.

02 Классификация с учетом типа задач, для решения которых используются данные

Для корректного выстраивания системы управления данными, помимо тех классификаций информации, которые представлены в 149-ФЗ, нужно учесть и другие. Так, данные можно классифицировать по типам задач, которые они решают. Например, выделяют данные для операционной деятельности и для аналитической.

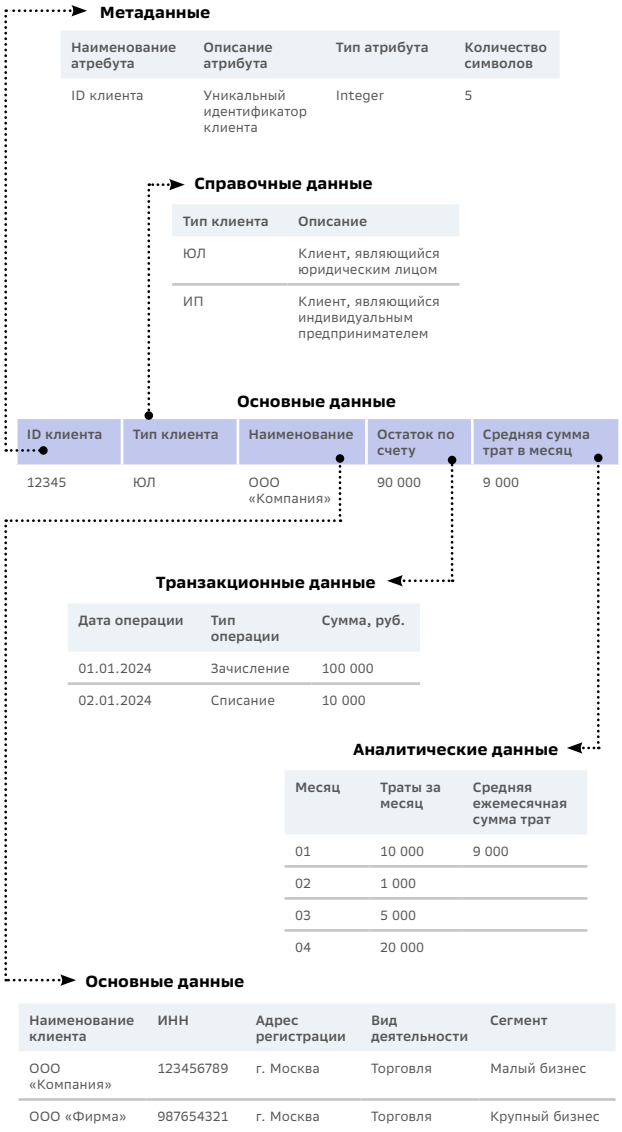
В выборку данных для операционной деятельности попадают:

- **метаданные** – это данные, описывающие структуру и характеристики данных (атрибуты, элементы);
- **справочные данные** – это данные из справочников, международных, общероссийских и отраслевых классификаторов и т. п.;
- **основные данные**[12] – это данные об объектах и бизнес-сущностях, представляющих ценность для организации, например, о наименовании клиентов, названии и свойствах продуктов, средствах и материалах и т. п.;
- **транзакционные данные** – данные, которые генерируются в результате выполнения конкретных транзакций или событий. Они представляют собой информацию о конкретных операциях (транзакциях), например, таких данных, как платежи, переводы, заказы и другие действия, которые происходят в режиме реального времени.

В выборку данных для аналитической деятельности попадают[13]:

- **аналитические данные** – это данные, являющиеся результатом систематизации и интерпретации информации.

В бизнес-процессах компании взаимосвязь видов данных можно отобразить следующим образом:



Как видно из схемы, для поддержания должного уровня качества, целостности и отказоустойчивости бизнес-процессов необходимо выстроить систему управления данными для каждого типа данных[14].

Помимо указанных данных, при построении системы управления данными необходимо учитывать, что данные не всегда находятся в базах данных в нужных ячейках. Зачастую данные изначально собираются компаниями в виде заполненных анкет или даже в свободном формате в виде рукописных заявлений. Такой вид данных требует особого внимания и ресурсов для управления. Например, сравним два вида реестров клиентов: один собран посредством АС, а другой представляет собой множество аудиозаписей телефонных разговоров, содержащих в себе информацию о клиентах. Очевидно, что удобство работы с данными для указанных случаев разное, и для второго случая требуется отдельная, достаточно трудоемкая, обработка данных, чтобы, например, создать базу данных адресов клиентов, продиктованных ими в ходе телефонного разговора. При этом обеспечение качества таких данных, а также поддержание их актуальности – затруднительно.

[12] Термин используется в ГОСТ Р ИСО 8000-2-2019 «Качество данных. Часть 2. Словарь», однако также часто используют термин «мастер-данные», описывающий то же самое.

[13] Важно отметить, что для целей аналитики зачастую используются также копии данных, изначально образовавшихся в операционной деятельности.

[14] Имеется в виду классификация данных с учетом типа задач, для которых они используются.

Существует три типа данных в зависимости от структуры:

Структурированные данные – это данные, имеющие строго фиксированную структуру, формат и существующие в фиксированном поле в рамках файла или записи. Также могут определяться моделью данных (логической или физической):

Реляционные БД

Таблицы

Полуструктурированные (слабоструктурированные) данные – это подвид структурированных данных, не имеющих строго определенной структуры или схемы базы данных, но предполагающих наличие правил, позволяющих выделять отдельные семантические элементы при их интерпретации – прежде всего правил расстановки тегов и других маркеров, отмечающих и выделяющих элементы данных. Такие данные могут менять свою структуру, наименования атрибутов, типы атрибутов:

Файлы CVS, xml, json

Yandex.Metrika, GAnalytics

Неструктурированные данные – это данные, произвольные по форме, не имеющие заранее заданной структуры или организации:

Текстовые документы

Фото, видео

Управление неструктурированными и полуструктурированными данными представляет собой наибольшую сложность. Представим, что нам нужно удалить персональные данные третьего лица, указанные клиентом в ходе телефонного разговора с банком, по факту получения банком требования этого лица о прекращении неправомерной обработки его персональных данных. Как это осуществить при необходимости сохранения полученных данных? Как оперативно узнать, какие данные, каких лиц и в каких разговорах содержатся? Это вопросы, ответы на которые необходимо найти при построении системы управления данными в этом примере.

03 Разнообразность инфраструктуры

Мы познакомились со всем разнообразием различных классификаций данных, которые могут встречаться в компаниях и которые необходимо учитывать при управлении данными, подсветили самые сложные вопросы и на примерах поняли необходимость управлять не только основными данными, например, о клиентах, но и метаданными, а также справочной информацией. Теперь перейдем от самих данных к местам их обработки, в том числе хранения. На первый взгляд кажется, что, перечислив рабочие места работников, автоматизированные системы и их базы данных, мы покроем всевозможные места обработки данных. Где еще они могут быть?

Начнем с работников: данные у них могут храниться на компьютерах и ноутбуках («толстый» клиент), на серверных частях («тонкий» клиент), а также на отчуждаемых носителях информации (например, USB-флешки, CD-диски) или находиться на общих файловых ресурсах, которые находятся в центре обработки данных компании или на внешних облачных ресурсах.

Говоря о центре обработки данных, также хотелось бы отметить несколько проблем:

- Существует несколько типов баз данных, принципиально различающихся по подходам к работе с данными. Самыми распространенными являются реляционные[15] базы данных. Однако для обеспечения деятельности крупных data-driven компаний, работы Data-science и ввиду специфики деятельности могут использоваться также документо-ориентированные базы данных, базы данных временных рядов, графовые базы данных, поисковые базы данных (Search Engines), объектно-ориентированные базы данных или векторные базы данных. Даже в рамках одного типа баз данных может использоваться множество различных СУБД разных производителей (разработчиков), например, Teradata, OracleSQL, Greenplum, MS SQL, PostgreSQL и др. Ввиду исторического развития инфраструктуры компании, использования автоматизированных систем различных разработчиков, одни и те же данные (их копии) могут находиться на витринах в разных типах баз данных, что усложняет процедуру построения горизонтального Data Lineage[16] для данных.
- Нельзя забывать и про наличие резервных копий баз данных. Если возникла ситуация и выяснилось, что действующая база данных более актуальна, чем резервная копия, то в случае сбоя и последующего восстановления из бэкапа, если компания не сделала бэкап после актуализации данных, под вопросом окажется, как минимум, актуальность данных, а в худшем случае – качество данных, и может возникнуть неправомерная обработка персональных данных, нарушение прав субъектов и риски привлечения к административной ответственности.

И конечно необходимо помнить про архивы, как электронные, так и архивы для бумажных носителей информации, обработка данных в которых регулируется специализированным законодательством[17].

При работе со всем множеством различных мест хранения данных могут возникать особые требования для конкретных видов данных. Так, например, для персональных данных актуальным является требование законодательства о невозможности совместного хранения и объединения баз данных, содержащих персональные данные, обрабатываемые в несовместимых целях.

04 Множество бизнес-процессов, использующих данные

Другим важным элементом в управлении данными является учет и анализ бизнес-процессов, использующих и порождающих данные. Ведь данные нужны не сами по себе, они используются для извлечения выгоды, оказания необходимых услуг, предоставления товаров или для соблюдения требований применимого законодательства. Со стороны каждого процесса могут выставляться свои требования к качеству, полноте и актуальности данных.

> Самым сложным в управлении данными является тот факт, что одни и те же данные могут использоваться разными процессами. Так, удовлетворяя требованиям одного процесса, мы можем не удовлетворять требованиям другого.

В рассмотренном выше примере, когда клиент обслуживается одновременно как ФЛ и ИП, можно наглядно увидеть, что данные этого клиента используются в разных не связанных между собой бизнес-процессах: одни и те же данные собираются разными клиентскими путями на различных правовых основаниях. Их обработка осуществляется в разных автоматизированных системах, в различных целях, с отличающимися сроками обработки. Применяя управляющее воздействие к данным, описанным в примере с ИП, необходимо убедиться, что это не противоречит иным бизнес-процессам. Так, например, при актуализации персональных данных необходимо произвести актуализацию во всех автоматизированных системах, а при уничтожении данных нужно быть уверенным, что уничтожаемые данные в одном бизнес-процессе не используются в других бизнес-процессах и их уничтожение не приведет к остановке бизнес-процессов или не нарушит связанности данных.

[15] Реляционная база данных – это составленная по реляционной модели база данных, в которой данные, занесенные в таблицы, имеют изначально заданные отношения.

[16] Горизонтальный Data Lineage – один из принципов в системе управления данными, предполагающий отслеживание перемещения данных от их источника до конечных точек, включая все промежуточные этапы трансформации и обработки данных. Значимость Data Lineage заключается в том, что реализация данного принципа позволяет отследить проблемы качества данных и другие ошибки до их первопричины и провести анализ влияния новых изменений на существующие объекты.

[17] Федеральный закон «Об архивном деле в Российской Федерации» от 22.10.2004 № 125-ФЗ.

05 Cost-to-Value и соответствие ожиданиям бенефициаров компании

Наверное, самый главный вопрос: а что будет, если ничего не делать? Ведь если мы говорим про коммерческую организацию, то ее основной задачей является получение прибыли. И тут возникает вопрос баланса трудозатрат на управление данными с одной стороны и возможностью поддержания выполнения бизнес-процессов на необходимом уровне с учетом возникающих рисков с другой стороны. Под рисками понимаются как риск данных и риск кибербезопасности, так и правовые и регуляторные риски.

Соответственно, перед построением системы управления данными в компании необходимо определить цели и потребность управления данными, то есть имеющиеся недостатки в текущей деятельности компании и ее бизнес-процессах, которые мы хотим закрыть, внедряя новые подходы в управлении данными. Может быть несколько базовых сценариев: минимизация рисков, устранение нарушений регуляторных требований или развитие бизнеса. В зависимости от целей выстраивается и подход: можно приоритизировать обеспечение безопасности и качества данных, а можно акцентировать внимание на аналитике, проверке гипотез и Data Science.

> Важно отметить, что внедрение системы управления данными, помимо выгоды для бизнеса, позволяет обеспечить необходимый уровень кибербезопасности данных на каждом этапе их жизненного цикла.

С чего начать управление данными

Итак, немного углубившись в суть и объем рассматриваемого вопроса, рассмотрим, наконец, с чего начать при построении системы управления данными. Подходов может быть множество, но могу порекомендовать рассмотреть следующий:

1. **Определяем три домена**, последовательная реализация которых позволит выстроить систему управления данными:
   - знания о данных;
   - знания о местах хранения данных, внутренних и внешних потоках данных;
   - непосредственно управление данными (определение применимых требований в соответствии с законодательством и требованиями кибербезопасности, соблюдение данных требований и обеспечение должного качества данных и правомерной их обработки) на всех этапах жизненного цикла.
2. **Определяем роли и ответственных.** Потребуется экспертиза подразделений информационной безопасности, ответственного за организацию обработки персональных данных (DPO[18]), ИТ-специалистов, владельцев данных и владельцев бизнес-процессов.
3. **Делаем первичный аудит** состава бизнес-процессов, типов данных и инфраструктуры компании. Определяем требования для каждого этапа жизненного цикла исходя из вида обрабатываемых данных. Создаем дорожную карту с определением задач, сроков и ответственных.
4. **Внедряем контрольные процедуры.** Никакие правила и организационные меры, выстроенные процессы и налаженные процедуры не могут полностью обходиться без процедур контроля. Будь то человеческий фактор или сбой в работе автоматизированных систем, необходимо своевременно выявлять отклонения и сбои в работе системы управления данными. Ведь бизнес не стоит на месте: появляются новые бизнес-процессы, новые виды данных, новые средства автоматизации и вместе с ними должна развиваться и система управления данными.

[18] DPO – Data Protection Officer (лицо, ответственное за организацию обработки персональных данных.

Выводы

В результате выстраивания системы управления данными удается достичь множества положительных результатов:

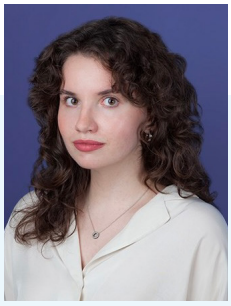
- Соответствие регуляторным требованиям – в случае с персональными данными достигается, например, за счет обработки персональных данных в законных и допустимых целях, объеме и сроках, прозрачности внешних потоков (с третьими лицами), контроля недопустимости ведения баз данных, содержащих персональные данные, обрабатываемые в несовместимых целях.
- Обеспечение защиты данных, включая персональные данные – предполагает обеспечение их конфиденциальности, целостности и доступности.
- Обеспечение приватности (Privacy) – защита права человека на неприкосновенность частной жизни и личной информации, предоставление возможности контролировать информацию о себе.
- Повышение качества данных – улучшение основных метрик, отвечающих за качество: полнота, своевременность, волатильность, точность, валидность, согласованность и наличие.
- Оптимизация мест хранения данных – сокращение количества входных точек при сборе/создании и последующего хранения данных, сокращение дублирования реплик и витрин, своевременное уничтожение ненужных данных.
- Развитие аналитики и Data Science – построение более глубокой аналитики, создание своих моделей искусственного интеллекта и повышение монетизации данных.

Таким образом, мы убедились, что культура работы с данными играет главную роль в деятельности data-driven компаний и позволяет достичь высоких финансовых результатов как за счет снижения рисков и издержек, так и за счет более грамотного извлечения прибыли из данных. Однако управление данными – это не самостоятельный процесс. Поэтому важно учитывать: он не может реализовываться в отрыве от бизнес-процессов компании, а наоборот, должен влиять и на сами бизнес-процессы, и на данные, которые в них обрабатываются, чтобы обеспечить требуемое качество и защиту, а также реализовывать обработку, минимизируя негативные последствия от возникающих рисков.

К СОДЕРЖАНИЮ

Управление персональными данными

в условиях непрерывного увеличения их объема



Яна Гришкова

Эксперт в области построения процессов кибербезопасности и приватности, команда DPO Блока Сеть продаж, Сбер

Екатерина Басниева

Эксперт в области приватности, дата-инженер, компания Группы Сбер

Вступление

В условиях высокой конкуренции, развивающейся в современном мире в различных сферах бизнеса, компании стараются своевременно выявить потребности клиентов и сократить возможные ошибки при формировании продуктовых предложений, так как каждая ошибка является упущенной выгодой для бизнеса. Для принятия решений, связанных с развитием продуктов и сервисов, предоставляемых конечным клиентам, компании стремятся к внедрению data-driven подхода.

По данным исследования, проведенного Gartner[1], к 2025 году 70% компаний будут использовать data-driven подход и начнут обгонять своих конкурентов за счет того, что внедрят подходы формирования дата-ориентированного управления предлагаемыми ими продуктами и сервисами.

В этой статье предлагаем поговорить о том, что такое data-driven подход, в чем разница между управлением данными (Data Management) и руководством данными (Data Governance), на что обратить внимание при внедрении функции управления данными, наконец, какова роль управления данными в рамках data-driven подхода в контексте регулирования персональных данных.

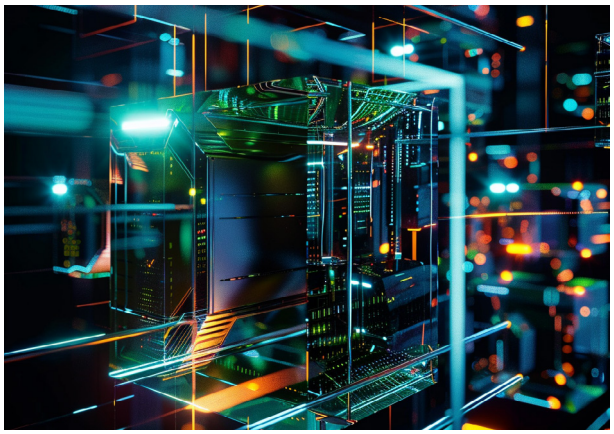
[1] Gartner. Gartner Research. Summary Translation: Predicts 2022: Data and Analytics Strategies Build Trust and Accelerate Decision Making. URL: https://www.gartner.com/en/documents/4009918 (дата обращения: 20.04.2024).

Что такое data-driven подход

Сама концепция data-driven (с англ. «управление на основе данных») подразумевает принятие решений об изменениях продуктов и сервисов на основе анализа накопленных данных на каждом этапе развития продукта. Такой подход позволяет найти узкие места в продуктах и оптимизировать их, используя зоны роста в качестве ключевых точек развития бизнеса, а значит преобразовать недостатки продуктов в дополнительную прибыль для бизнеса.

Внедрение data-driven подхода предполагает:

- итеративный сбор и последующее накопление больших объемов данных, включающих знания о взаимодействии конечных пользователей с предоставляемыми компанией продуктами, сервисами;
- изучение поведения пользователей, их предпочтений, что важно для создания и развития продуктов, сервисов, которые способны удовлетворять запросы пользователей и привлекать новых клиентов (на это мы просим обратить особое внимание в контексте тематики персональных данных);
- тестирование гипотез – очевидно, что для того, чтобы понять, какие изменения в продуктах будут способствовать улучшению различных метрик эффективности (в частности, метрик привлечения, вовлеченности, производительности, финансовых метрик), нужно проводить аналитику и тестировать различные гипотезы (например, о том, какой пользовательский интерфейс наиболее удобен клиентам и приводит к результату, который ожидает бизнес – будь то совершение покупки или любого иного действия, в котором бизнес измеряет эффективность продукта или сервиса);
- командную работу заинтересованных подразделений компании, каждое из которых нацелено на получение результата.



Жизненный цикл data-driven подхода включает шесть этапов:

1. **Идея**, когда появляется гипотеза о необходимости развития продукта, сервиса.
2. **Проверка гипотезы**, когда оценивается ее состоятельность.
3. **Тестирование гипотезы**, то есть усовершенствование текущего или запуск нового продукта, сервиса, если гипотеза оправдала себя (то есть если выявлена необходимость в изменении продукта, сервиса или запуске нового с учетом потребности пользователей).
4. **Наблюдение за эффектом** и влиянием произошедших изменений в виде усовершенствованного продукта, сервиса или нового запуска.
5. **Формирование выводов** о влиянии изменений на пользователей (удовлетворили ли внедренные изменения, требуются ли дополнительные улучшения, если да, то какие).
6. **Проектирование новых гипотез** с учетом сделанных выводов (например, получив положительный эффект от реализованных улучшений продукта или сервиса, бизнес может сформировать гипотезу о необходимости развития проведенных улучшений по ранее определенному вектору).

Итого, мы уже можем сделать вывод о том, что data-driven – это подход, в основе которого лежит регулярное наблюдение за потребностями пользователей, соответствующая аналитика и интерпретация этих потребностей для их последующей реализации в качестве улучшений для продуктов и сервисов.

По сути такой подход позволяет найти «узкие» места в продуктах и сервисах (то есть те, которые могли бы быть не замечены при верхнеуровневом анализе) и оптимизировать их за счет:

- выявления возникающих потребностей пользователей продуктов и сервисов;
- усовершенствования текущего функционала, предоставляемого пользователю в случае, если он не вполне удовлетворяет имеющимся запросам или является неудобным для конечного пользователя;
- принятия решений об отказе от функционала, который препятствует эффективному взаимодействию пользователя и сервиса;
- определения связи между регионами и предлагаемыми продуктами и сервисами (очевидно, что, запуская продукт или сервис, нужно понимать, возымеет ли это положительный эффект в том или ином регионе);
- определения наиболее выгодных вариантов маркетинговых коммуникаций.

Таким образом, ключевое преимущество data-driven подхода состоит для бизнеса в том, что, используя зоны роста в качестве ключевых точек развития, можно преобразовать недостатки продуктов и сервисов в достоинства и, соответственно, в дополнительную прибыль и лояльность клиентов.

У data-driven подхода, помимо описанных преимуществ, впрочем, есть и недостатки.

Во-первых, использование такого решения – недешевое удовольствие, потому что для его качественной работы требуется много ресурсов, включая человеческие (наем работников с необходимыми компетенциями) и финансовые ресурсы (в связи с трудоемкостью внедрения процессов и технологий).

Во-вторых, работать с данными не так просто, потому что необходимо регулярно обеспечивать и поддерживать:

- необходимое качество данных;
- доступность требующихся для анализа данных в конкретный момент времени;
- надежность данных, то есть их точность, полноту и непротиворечивость.

При этом всегда есть риск ошибок в аналитике данных (например, из-за неверного толкования гипотезы или попадания в ловушку ненадежных данных, что возможно, если компания закупает данные из различных источников) или нарушений контекста (случаи, когда на момент прогнозирования развития продукта не было учтено внешнее влияние на продукт, которое не зависит от условий его предоставления).

Тем не менее, несмотря на то, что рассматриваемый нами подход принятия решений не лишен недостатков (и, если быть честными, преодоление недостатков требует существенных ресурсов), он по-прежнему является очень востребованным. Во многом это связано с его ориентированностью на потребности клиента, которая достаточно закономерна: с учетом технологического роста и имеющихся у компаний накопленных данных принятие решений «вслепую» является непозволительной роскошью, влекущей существенные риски (например, репутационные и финансовые), которые могут возникнуть в случае неправильно принятых решений.

Поэтому бизнес заинтересован в:

- создании условий для повышения качества данных, их доступности и надежности;
- развитии инструментов по управлению накопленными данными;
- наращивании экспертизы и квалификации работников, которые занимаются аналитикой, тестируют гипотезы, принимают решения.

Помимо этого, не на последнем месте у бизнеса стоит вопрос обеспечения соответствия деятельности, которая осуществляется в рамках data-driven подхода, применимому законодательству.

Как становится понятно из вышеописанной информации, наибольшее влияние на принятие решений оказывают данные конечных пользователей. А любые данные, позволяющие прямо или косвенно установить личность пользователя (субъекта-физического лица), являются персональными данными. Поэтому в скоуп законодательных требований, которые нужно выполнять бизнесу, попадают требования законодательства о персональных данных. Обработка персональных данных клиентов налагает на компании законодательные обязательства, в том числе по выстраиванию системы организации обработки и защиты персональных данных в соответствии с требованиями законодательства, а также необходимость в управлении данными на этапах их жизненного цикла, обеспечивая, в частности, правомерность обработки персональных данных, их актуальность и своевременное уничтожение. Причем это важно не только с точки зрения предотвращения правовых и регуляторных рисков (что особенно актуально на фоне законопроектов об оборотных штрафах за утечки персональных данных), но и во многом с позиции исключения репутационных рисков, повышения лояльности клиентов за счет выстраивания отношений «бизнес-клиент», которые позволят клиенту быть уверенным в том, что передавать данные в ту или иную компанию надежно и безопасно.

Итого у бизнеса две крупные задачи:

- создать условия для повышения качества данных, их доступности, надежности и сформировать инструмент по управлению обрабатываемыми данными;
- исключить правовые, регуляторные, репутационные риски за счет соблюдения требований применимого законодательства, в частности, законодательства о персональных данных.

В случае с обработкой данных – особенно персональных данных, тем более в компании, где много клиентов и есть еще другие субъекты персональных данных (например, работники) – второе без первого невозможно.

Поэтому предлагаем разобраться, что нужно делать для того, чтобы все это реализовать и в конечном счете заставить данные приносить прибыль. В первую очередь давайте начнем с основ и поймем, что такое Data Governance (Руководство данными) и Data Management (Управление данными на основании Руководства) и в чем их отличия.

Data Governance

Руководство данными (от англ. Data Governance) – это концепция стратегического управления данными, которая определяет данные как ключевой актив организации, требующий внедрения решений для управления данными как ценным активом, что соответствует идее data-driven подхода.

Необходимость внедрения подхода по руководству данными сформировалась исходя из перехода человека в онлайн-среду: нам больше не нужно выходить из дома, чтобы перевести деньги со счета на счет, жители крупных городов не представляют свою жизнь без доставки товаров на дом. Множество функций, предоставляемых конечным покупателям в интернете, рождают большое количество данных, а значит и знаний о человеке, его привычках и потребностях. При ежедневном увеличении спроса на сервисы и продукты, предоставляемые посредством сети интернет, количество обрабатываемых данных в компаниях также многократно увеличивается. Наиболее заметно это увеличение для банковской отрасли и e-commerce (от англ. электронная коммерция).

Внедрение подхода по руководству данными гарантирует возможность управления массивами данных для их последующего эффективного использования в различных целях: при проведении аналитики данных, построении AI-моделей, формировании отчетности, внедрении автоматизации ручных процессов и многих других целей, которые перед собой может поставить бизнес (конечно, сделаем оговорку: для всего этого нужно, как минимум, обеспечить правовые основания обработки персональных данных, используемых в тех или иных целях).

В настоящее время концепция Data Governance ввиду ее комплексности охватывает планирование, мониторинг и обеспечение реализации соответствующих мер управления накопленными в компании данными, позволяя достигнуть сразу нескольких целей, включая:

- повышение надежности данных;
- снижение затрат на хранение данных;
- обеспечение безопасности данных;
- соответствие требованиям применимого законодательства.

Data Governance vs Data Management

Понятия Руководство данными (Data Governance) и Управление данными (Data Management) зачастую не разделяют, хотя это сущности, которые расположены на разных уровнях принятия управленческих решений.

Управление данными, в отличие от Руководства данными, представляет процесс, который включает в себя разработку, реализацию и контроль соблюдения политик, программ и практик, направленных на предоставление, проверку, защиту и повышение ценности данных в течение их жизненного цикла.

Управление данными включает в себя несколько функций, среди которых выделяется руководство данными. Однако руководство данными занимает главенствующее положение по отношению к другим функциям, поскольку оно на высшем уровне определяет стратегию управления данными[2].

В отличие от управления данными, которое сосредоточено на извлечении ценности из данных как информационных активов, руководство данными направлено на принятие решений относительно данных и организацию работы людей и процессов, связанных с данными.

Кто? Как? С помощью чего?

Сложно представить, что data-driven подход можно реализовать без эффективной реализации функции управления данными.

С чего начать реализацию? Ответить на вопросы: Кто? Как? С помощью чего?

В основе руководства данными лежат следующие компоненты функции управления данными: люди, процессы, технологии.

Развитие каждого из перечисленных компонентов необходимо закладывать при формировании дорожной карты по совершенствованию функции управления данными в компании. Разберем их более подробно.

[2] Подробнее: https://atlan.com/data-governance-vs-data-management/ (дата обращения: 13.05.2024).

Процессы

Эффективный data-driven подход невозможен без этапа выстраивания и дальнейшего поддержания процессов управления данными. Выстраивание процессов необходимо начинать с формирования верхнеуровневого представления о том, какие цели и задачи ставит перед собой функция. На эти вопросы может ответить сформированная в компании стратегия управления данными.

Стратегия управления данными, как и любая другая стратегия, должна соответствовать бизнес-стратегии компании, что означает ориентацию на достижение целей бизнеса за счет внедрения процессов управления данными. При разработке стратегии необходимо учитывать контекст компании, включая масштаб компании, уровень зрелости ИТ-процессов, процессов кибербезопасности и приватности, объемы обрабатываемых данных и задачи, для которых используются данные.

Для корректного определения контекста компании следует провести комплексный аудит, по результатам которого сформируется понимание текущего состояния процессов компании и определение целевого видения результата внедрения процесса управления данными.

Для определения направлений деятельности по выстраиванию функции управления данными следует руководствоваться лучшими практиками в рассматриваемой области. В настоящий момент наиболее известны два фреймворка:

- свод знаний по управлению данными (Data Management Body of Knowledge, DMBoK), разработанный Международной ассоциацией управления администрирования данных (Data Administration Management Association, DAMA);
- модель оценки способностей по управлению данными (Data Management Capability Assessment Model, DCAM) Совета по управлению корпоративными данными (Enterprise Data Management Council, EDM Council).

Основное различие между подходами кроется в самом названии фреймворков:

- модель оценки способностей по управлению данными представляет собой руководство по проведению комплексной оценки зрелости функции управления данными и поиску возможностей по улучшению функции;
- свод знаний по управлению данными является комплексным руководством, раскрывающим принципы и содержащим практики в рассматриваемой области.

Более того, Свод знаний по управлению данными предназначен для более широкой аудитории и охватывает большее количество направлений управления данными, включая обеспечение безопасности и приватности обрабатываемых данных.

Однако вне зависимости от выбранного фреймворка необходимо помнить о том, что применимое законодательство в области приватности накладывает дополнительные требования на обработку и защиту данных, в соответствии с чем при планировании изменений, затрагивающих обработку данных в компании, необходимо оценить требования законодательства и учесть их во избежание возможных рисков, реализация которых может повлечь как убытки (в виде упущенной выгоды или реального ущерба), так и репутационный вред.

Для внедрения процессов управления данными, вне зависимости от выбранных фреймворков, необходимо предусмотреть несколько основных составляющих, которые позволят эффективно внедрить и контролировать процесс:

- **терминология** – документирование терминологии, используемой компанией в процессах управления данными, так как в настоящее время отсутствует стандартизированная и устоявшаяся терминология в рассматриваемой области;
- **описание процесса** – документирование информации о сущности процесса и его порядке выполнения, необходимое для обеспечения повторяемости процесса;
- **ответственность** – закрепление ответственности за внедрение процесса и его последующую реализацию в зависимости от исполняемой роли участника процесса, а также меры воздействия, реализующиеся в случае выявления нарушений в процессе;
- **контроль** – определение порядка проведения контрольных мероприятий, в том числе: ответственности, методики, частоты и требований к документированию результатов проведения таких мероприятий;
- **пересмотр** – фиксация случаев необходимости пересмотра процесса и ответственности за пересмотр.

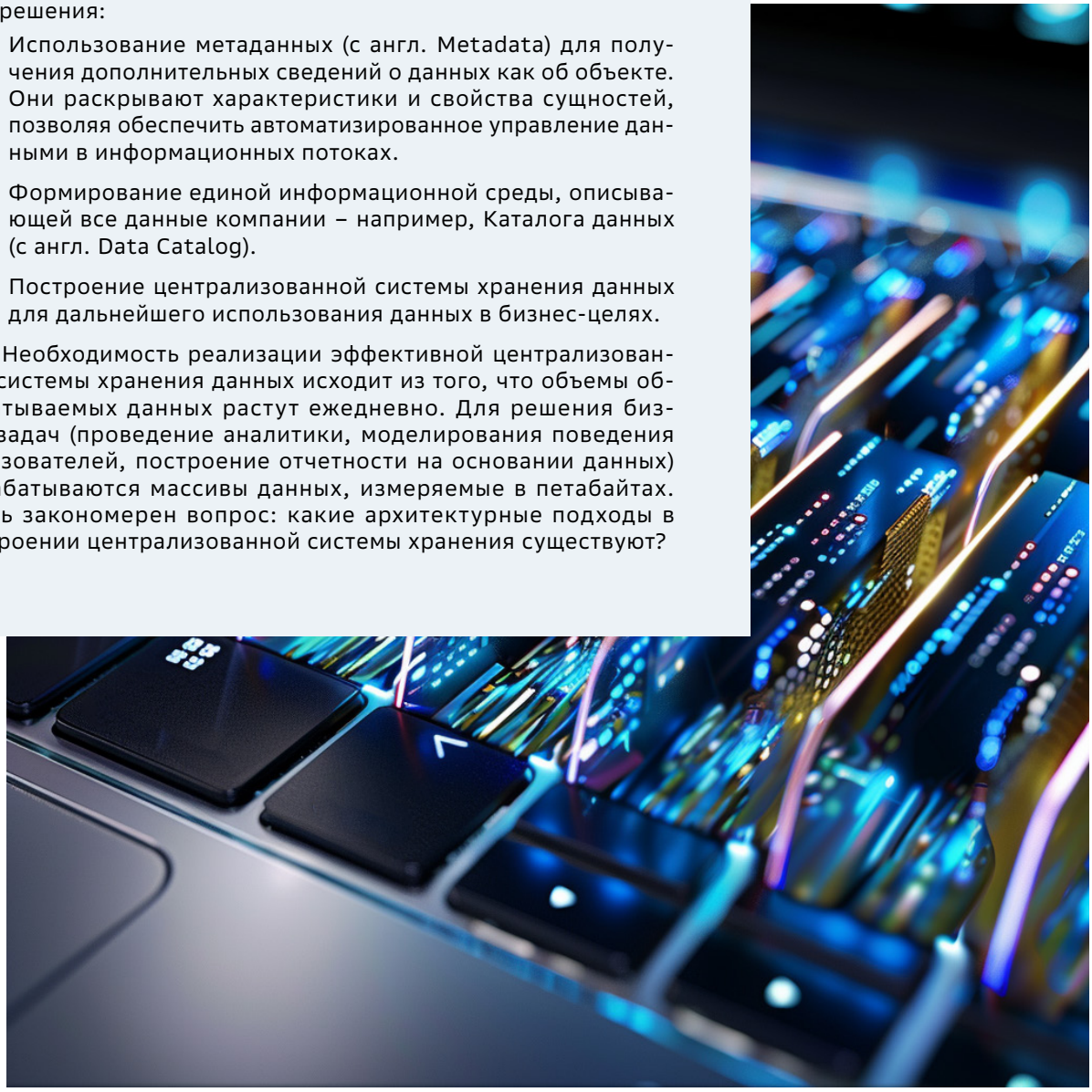
Технологии

Обработка данных в современном мире неразрывно связана с информационными технологиями, доступ к данным происходит через взаимодействие работников с информационными системами. Потребность в многообразии информационных систем для выполнения рабочих задач различных бизнес-подразделений ведет к увеличению используемых в компаниях информационных систем, и, как следствие, многообразию хранилищ данных. Данная ситуация становится причиной дублирования данных, отсутствия согласованности данных между системами, неудобствами для клиента (вспомните ситуацию, когда один и тот же человек, желая заменить документы, вынужден обращаться в компанию дважды – один раз как физическое лицо, а в другой – как физическое лицо со статусом индивидуального предпринимателя).

Для решения данной проблемы бизнес реализует следующие решения:

- Использование метаданных (с англ. Metadata) для получения дополнительных сведений о данных как об объекте. Они раскрывают характеристики и свойства сущностей, позволяя обеспечить автоматизированное управление данными в информационных потоках.
- Формирование единой информационной среды, описывающей все данные компании – например, Каталога данных (с англ. Data Catalog).
- Построение централизованной системы хранения данных для дальнейшего использования данных в бизнес-целях.

Необходимость реализации эффективной централизованной системы хранения данных исходит из того, что объемы обрабатываемых данных растут ежедневно. Для решения бизнес-задач (проведение аналитики, моделирования поведения пользователей, построение отчетности на основании данных) обрабатываются массивы данных, измеряемые в петабайтах. Здесь закономерен вопрос: какие архитектурные подходы в построении централизованной системы хранения существуют?



Архитектурные подходы

В настоящее время широко известны несколько архитектурных подходов в проектировании централизованной системы хранения данных, разница в которых проявляется в определении конечной структуры хранимых данных, а следовательно, и применяемым к данным преобразованиям:

Хранилище данных (от англ. Data Warehouse) – хранилище, в котором заранее определена структура хранимых данных, то есть атрибуты данных заранее определены. Для поддержания структуры хранилища данные проходят ETL-преобразования (от англ. Extract, Transform, Load), то есть данные извлекаются из исходных систем, преобразуются в нужный формат и загружаются в хранилище данных. То есть под хранилищем данных в данном случае понимается хранилище реляционных данных. Основное отличие от реляционных баз данных в таком случае – это масштаб хранимых данных.

Озеро данных (от англ. Data Lake) – хранилище, содержащее данные, формат которых не определен заранее. В данном случае применяется подход ELT-преобразований (от англ. Extract, Load, Transform), то есть данные загружаются в хранилище в исходном виде для последующего проведения необходимых преобразований. То есть в данном случае обеспечивается хранение копии неструктурированных исходных данных, которые в дальнейшем могут быть преобразованы для целей создания реляционного хранилища данных. Выбор данной структуры позволяет обеспечить гибкость хранения данных[3].

Однако в условиях современных задач компании используют комбинированных подход, предполагающий двухуровневую архитектуру централизованной системы хранения данных: данные из источников попадают в озеро данных (Data Lake) для их дальнейшего приведения к структурированному виду и загрузки в хранилище данных (Data Warehouse). Эта концепция носит название Lake Warehouse. Схематично ее можно изобразить так:

Двухуровневый подход (Lake Warehouse) к формированию хранилища является оптимальным по временным затратам на доступ к данным, а также позволяет обеспечить целостность данных ввиду организации промежуточного хранилища данных, однако он является более дорогим по сравнению с использованием каждого из ранее описанных подходов в отдельности.

После определения архитектуры корпоративного хранилища данных необходимо предусмотреть **возможность быстрого поиска данных и отслеживания потоков данных**.

[3] Gartner. https://www.gartner.com/en/information-technology/glossary?startsWith=D (дата обращения: 13.05.2024).

Поиск данных по метатегам

Метаданные играют ключевую роль в управлении данными, обеспечивая их эффективное использование. Они представляют собой информацию об используемых в компании данных, которая содержит их контекст. Различают несколько типов метаданных:

- описательные метаданные (например, описание содержания);
- административные метаданные (информация об источнике, дате получения, уровне критичности данных и правах доступа);
- структурные метаданные (описание взаимосвязей между данными).

Для обеспечения возможности управления данными используются инструменты разметки данных – данные помечаются тегами, что позволяет получить дополнительные сведения о данных. Например, о категории обрабатываемых данных (применительно к персональным данным это может быть тег, свидетельствующий о категории конфиденциальности данных или категории персональных данных: биометрические, специальные, иные), владельце данных, в случае обработки персональных данных – также о правовом основании на обработку (например, сведения о дате получения согласия на обработку персональных данных или дате заключения договора, которые являются основанием обработки персональной информации).

Отслеживание потоков данных

Отслеживание потоков данных внутри компании на каждом из этапов жизненного цикла данных, включая различные модификации данных, возникшие в процессах их обработки, возможно с применением функций отслеживания пути данных (от англ. Data Lineage). Использование инструментов отслеживания данных позволяет значительно повысить качество данных и обеспечить надежность используемых данных, так как с их помощью можно отследить источники данных и дальнейшие способы их обработки.

> В совокупности использование метаданных и инструментов отслеживания данных (Data Lineage) позволяет получить исчерпывающую информацию о происхождении, изменении и обработке данных.

Каталог данных

Реализация задачи структурирования данных и облегчения их поиска среди массивов обрабатываемых данных выполняется с использованием такой сущности как каталог данных. Для формирования каталога данных необходимо определить все источники данных компании (информационные системы), а также атрибуты данных, обрабатываемые в рамках информационных систем. Каталог данных фиксирует метаданные, позволяя хранить и отображать информацию о данных конечным пользователям.

Задачами управления данными с точки зрения используемых технологий занимаются работники IT-подразделений компаний во взаимодействии со стюардами данных, которые отвечают за внедрение процессов управления данными.

После того, как мы разобрались с тем, как работают процессы и технологии, необходимо определить ответственность за их реализацию.

Люди

Являясь главным активом компании, работники взаимодействуют с данными ежедневно, что делает их участниками процесса управления данными. При этом каждый из них выполняет различные роли в процессах управления данными.

Роли в процессах управления данными зачастую связаны с ролевой моделью управления доступом. В самом деле, выполнение различных трудовых обязанностей требует обеспечения различного уровня доступа к данным. Разграничение доступа к данным – ключевой элемент обеспечения безопасности персональных данных и их конфиденциальности с учетом принципа предоставления доступа к персональным данным в минимальном объеме, необходимом для исполнения трудовых обязанностей.

Для того, чтобы оптимизировать и упростить процесс управления доступом к данными, необходимо определить конечный перечень ролей в процессе управления данными, зависящий от функциональных обязанностей работников компании. Ролевая модель позволяет назначать пользователям определенные роли, каждая из которых включает в себя набор прав и доступов, необходимых для выполнения конкретных задач или функций.

Для упрощения понимания **построения ролевой модели** можно выделить следующие группы и роли:

- Производители данных – бизнес-подразделения, реализующие конечный функционал сервиса, предоставляемый клиенту и «генерирующий» за счет этого входной поток данных в компанию.
- Пользователи данных – все конечные потребители данных, в чьи обязанности входит моделирование и исследование данных.
- IT-специалисты – проектировщики модели хранилища данных, разработчики хранилища данных, DB-инженеры, системные аналитики.
- Участники функции руководства данными – участники процесса управления с функциональными обязанностями в области руководства данными.

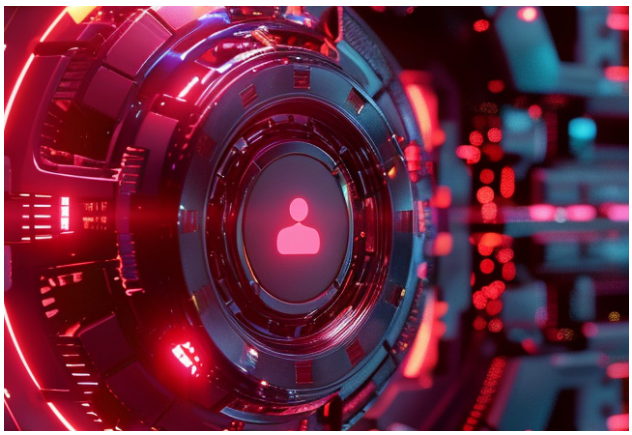
В рамках фреймворка DMBoK для функции руководства данными выделяются следующие ключевые роли ответственных:

- Руководитель цифровой трансформации (от англ. Chief Digital Transformation Officer, CDTO) – ответственный за внедрение процессов и технологий, которые связаны с цифровизацией и цифровой трансформацией (обновление процессов за счет внедрения цифровых решений).
- Руководитель по работе с данными (от англ. Chief Data Officer, CDO) – ответственный за внедрение процессов и технологий, необходимых для обеспечения работникам компании быстрого и безопасного доступа к данным в компании.
- Специалист по описанию данных, стюард данных (Data Steward) – ответственный за полноту и качество описания данных (метаданных), а также за внедрение и контроль процессов управления данными.
- Владелец данных – владелец сервиса или бизнес-процесса, в рамках которых в компании генерируются данные, за качество которых и за доступ к которым несет ответственность владелец.

Вместе с тем, в контексте **управления персональными данными** важно еще выделить следующие роли:

- Ответственный за организацию обработки персональных данных (от англ. Data protection officer, DPO) – ответственный за организацию обработки и определение требований к защите персональных данных в компании, обеспечение соответствия ее деятельности требованиям применимого законодательства в области обработки и защиты персональных данных.
- Ответственный за информационную безопасность (от англ. Chief Information Security Office, CISO) – ответственный за обеспечение кибербезопасности (в том числе безопасности персональных данных).

Часть ролей уже может присутствовать в компании, которая стоит у истоков внедрения процессов управления данными, однако следует помнить о том, что новые роли накладывают новый функционал на работников компании. При создании ролевой модели управления данными необходимо определить круг необходимых ролей, а также цели, задачи и необходимую квалификацию работников каждой роли.



Как функция управления данными влияет на выполнение требований законодательства в области персональных данных?

Для обеспечения контроля и надлежащей обработки данных субъектов персональных данных (физических лиц, чьи данные обрабатываются в компании) во многих развитых странах реализуется законодательный контроль за обработкой персональных данных.

Так, обработка персональных данных в Российской Федерации должна осуществляться в соответствии с требованиями Федерального закона «О персональных данных» от 27.07.2006 № 152-ФЗ (далее – 152-ФЗ), что предполагает выстраивание процессов, в которых осуществляется обработка с учетом установленных законодательных норм в части принципов, целей, условий, состава данных, сроков обработки, требований к обеспечению безопасности данных и многого другого.

Ранее мы уже отметили, что одной из задач бизнеса при использовании data-driven подхода является соблюдение требований законодательства о персональных данных. Кроме того, мы рассмотрели, какое влияние на соблюдение законодательных требований оказывает применение Data Governance и Data Management.

Вполне очевидно, что без внедрения в бизнес-процессы процессов управления данными обеспечить соответствие обработки персональных данных требованиям 152-ФЗ будет крайне затруднительно, особенно, если компания обрабатывает существенные объемы персональной информации во множестве бизнес-процессов, для каждого из которых применимы «свои» основания обработки, цели, категории и перечень данных, категории субъектов, чьи данные обрабатываются, допустимые сроки обработки и не только.

Теперь предлагаем более детально, на примерах, рассмотреть плюсы внедрения процессов управления данными при реализации требований применимого законодательства о персональных данных.

Одним из плюсов является возможность сформировать **реестр обработки персональных данных**. Реестр – это необязательное требование законодательства, однако это то, что значительно упростит процесс управления персональными данными в компании.

В случае реализации разметки обрабатываемых данных формирование реестра обработки персональных данных сводится к автоматизированной выгрузке атрибутов обрабатываемых данных в зависимости от целей обработки персональных данных из соответствующих автоматизированных систем. Реестр может содержать различную информацию, включая:

- наименование бизнес-процесса (цель обработки данных);
- подразделение-владелец бизнес-процесса;
- ответственный от владельца бизнес-процесса;
- источник получения персональных данных (субъект персональных данных, представитель субъекта, государственные или муниципальные органы, другое подразделение внутри компании);
- цели сбора и последующей обработки персональных данных;
- правовые основания сбора и последующей обработки персональных данных;
- категории субъектов персональных данных;
- категории и перечень персональных данных;
- цели передачи персональных данных;
- основания передачи персональных данных;
- автоматизированные системы, в которых хранятся данные, и т.д.

Причем между атрибутами можно провести связи, требующиеся согласно 152-ФЗ: например, можно соотнести цели и правовые основания, цели и состав данных, цели сбора и передачи и т.д. Таким образом, компания получит наглядную картину о том, насколько обработка персональных данных в процессах соответствует требованиям 152-ФЗ, и возможность обеспечить контроль за их правомерной обработкой.

Еще один плюс – **реализация права субъекта на доступ к его персональным данным**. Представьте ситуацию: физическое лицо является клиентом компании А и пользуется определенным спектром продуктов и услуг. Для этого у клиента заключено несколько договоров, подписаны согласия на обработку персональных данных. По каждому продукту и услуге обрабатывается разный объем данных, в различных целях. В определенный момент клиент решил запросить информацию о том, какие его данные обрабатывает компания, на каких основаниях и в каких целях по каждому имеющемуся у него продукту и услуге. Очевидно, что без внедрения в бизнес-процессы процессов управления данными предоставить клиенту такую информацию, тем более оперативно, невозможно. Реализация функции управления данными помогает автоматизировать выполнение реализации рассматриваемого права субъекта путем выгрузки данных из систем компании.

Помимо этого, среди преимуществ – **возможность отслеживания использования чувствительных категорий персональных данных субъектов** (специальные категории персональных данных, биометрические персональные данные).

В случае реализации разметки персональных данных и тегирования чувствительных категорий персональных данных можно ограничить возможность использования таких данных исходя из целей обработки.

Кроме того, плюсом является в**озможность отслеживания передачи персональных данных третьим лицам, включая трансграничную передачу**. Инструменты отслеживания данных позволяют реализовать отслеживание и контроль соответствия состава передаваемых персональных данных правовым основаниям на их обработку, включая отслеживание исходящего потока на территорию другого государства.

Наконец, в качестве плюса можно привести условия для обеспечения необходимого уровня защиты данных в зависимости от их категории. Законодательные требования накладывают на компании обязанности по реализации дополнительных мер защиты персональных данных с учетом уровня их чувствительности. Разметка обрабатываемых данных помогает определить требования по внедрению мер и средств защиты данных в случае обработки данных в рамках рассматриваемой информационной системы.

Заключение

Реализация data-driven подхода невозможна без внедрения функции управления данными.

Владение инструментами управления данными – это возможность оставаться конкурентоспособной компанией, генерирующей выручку и имеющей лояльных клиентов. Данные без реализации возможности управления ими из ценного актива компании превращаются в накапливаемый годами балласт.

В общем случае можно выделить следующие выгоды реализации функции управления данными для бизнеса:

1. Исключение рисков несоответствия законодательным требованиям в рамках процессов обработки данных и нарушения прав субъектов персональных данных.
2. Снижение Time-to-market продуктов, сервисов, услуг бизнеса за счет своевременного доступа к данным в процессах аналитики.
3. Повышение качества предоставляемых продуктов, сервисов, услуг за счет увеличения точности решений, принимаемых на основании надежных данных, используемых для анализа.
4. Разграничение ответственности в процессах обработки данных в компании.
5. Оптимизация конечной стоимости обработки данных в компании исходя из повышения скорости доступа к данным и выбора оптимальной архитектуры хранилища данных.

Однако не стоит забывать о том, что во главе развития продукта лежат потребности клиентов, представление о которых складывается у бизнеса из обрабатываемой информации об их персональных предпочтениях, действиях на сайте и в приложении, оценок, которые они ставят продуктам, услугам, сервисам, регулярности их использования и других сведений, которые неразрывно связаны с клиентом (определяют его или позволяют определить). Напрявляя фокус внимания на обработку персональных данных клиентов, важно понимать: обработка такой информации будет сопровождаться необходимостью реализации требований законодательства в области персональных данных.

Внедрение процессов управления данными является трудоемким и долгим процессом. Однако реализация комплексной системы управления данными, охватывающей управление персоналом, процессами и технологиями, может значительно повысить эффективность работы компании за счет обеспечения своевременного доступа к необходимым данным для решения бизнес-задач и снизить риски нарушения требований законодательства в области обработки и защиты персональных данных.

К СОДЕРЖАНИЮ

Графовые технологии

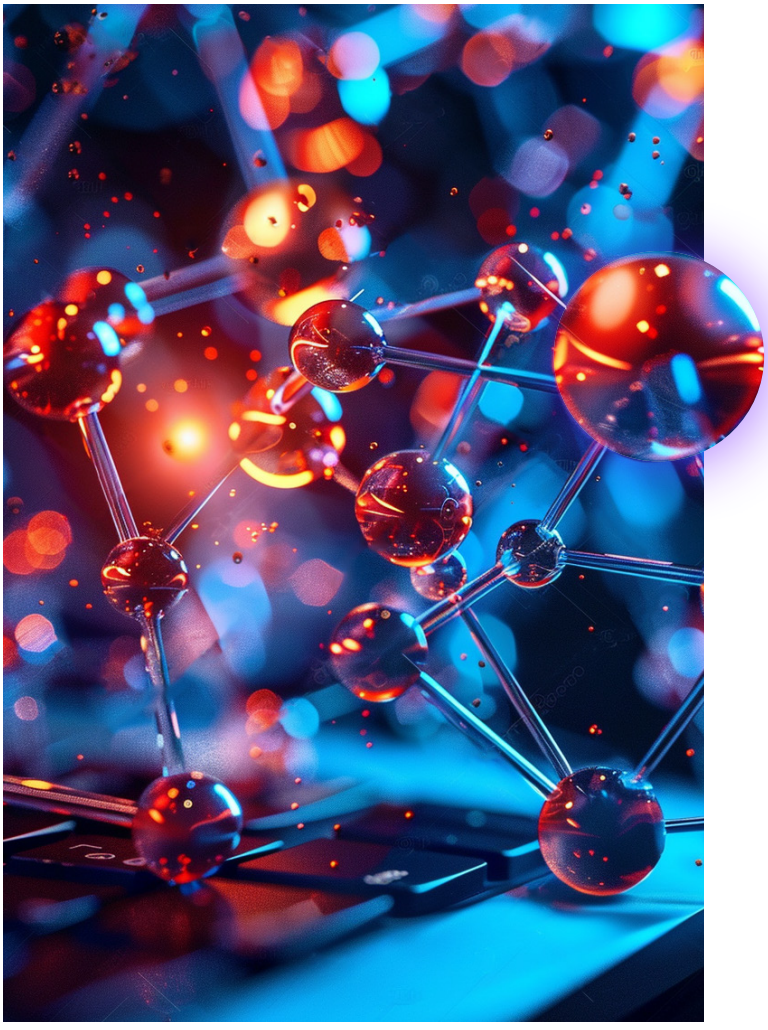
как эффективный способ управления данными

Алексей Булавин

Исполнительный директор управления развития технологий искусственного интеллекта и машинного обучения, SberData

Вступление

В современных реалиях, когда технологии работы с данными имеют первоочередное и зачастую определяющее значение для дальнейшего выстраивания процессов и операций, появление различных способов аналитики данных представляется неизбежным. Сфера управления данными при этом не стала исключением: работа с традиционными инструментами управления уже не достигала желаемых показателей эффективности, в особенности, если объем данных был очень большим. Появление графовой аналитики, способствующей изучению парных отношений между объектами, обусловило реализацию возможности принимать стратегические решения, в том числе увеличивающие конкурентные преимущества компаний, которые могут быть направлены, в частности, на обеспечение кибербезопасности и защиту персональных данных. В этой статье предлагаем поговорить о том, что такое графовые технологии, какую роль в управлении данными они играют, как используются для решения задач по защите данных, и как эти технологии развивает Сбер.



Что такое графы: немного базовой информации

Как вы уже успели догадаться по вступлению к статье, граф – это не только дворянский титул. Это еще некая математическая абстракция реальной системы окружающего нас мира, объекты которого обладают парными связями. Как математическая модель графы представляют собой совокупность узлов и ребер (связей) между ними. Представление имеющихся данных в виде графов помогает увидеть структуру данных визуально, применить те или иные идеи, а также определить возможные пути и способы разрешения поставленных задач.

Изначально графы рассчитывали на компьютерах исключительно для исследовательских или научных целей. Однако позже стало понятно, что графы можно применять и в бизнесе, где большой объем данных позволяет найти неочевидные связи, логические цепочки и использовать эту информацию для улучшения эффективности существующих или развития новых бизнес-инициатив.

Получилось это, правда, не сразу. Компьютеры для обработки графов требовали больших вычислительных ресурсов, а на тех мощностях, что были доступны, скорость была низкой. И если в науке зачастую можно не торопиться и подождать, когда вычисления завершатся, то в бизнесе требуются только быстрые решения.

Первые шаги

Изначально хранение графов и работа по поиску связей велись напрямую с реляционными базами данных через «select», до боли знакомый многим программистам. Серьезные сложности при такой архитектуре начинались уже на этапе построения относительно небольших графов, не говоря уже обо всем объеме данных, на котором нужно выполнять сложные графовые расчеты.

Такую задачу команда Сбера решала в 2016 году. Тогда было очевидно, что графовые технологии пока не готовы выполнять даже элементарные задачи с учетом объемов данных Сбера, среди которых в том числе персональные данные более 100 млн клиентов. Тем не менее, гонка уже началась: графовые технологии активно тестировали игроки уровня Google, поэтому задачей было – внедрить графовые технологии в процессы так, чтобы с их помощью среди прочего можно было выполнять задачи по эффективному управлению большими объемами данных.

> Решением стали разработка и последующее внедрение в бизнес-процессы Сбера собственных технологий обработки данных, представленных в виде больших графов.

В основу технологий будущей графовой платформы Сбера[1] было заложено новое технологическое ядро – собственная графовая база данных. Важно было предусмотреть возможность решения задач как трудоемкого пакетного расчета, так и легких онлайн-задач. Кроме этого, необходимо было разделять информацию на ту, которую необходимо держать и обрабатывать непосредственно в графе, и информацию, связанную с ним и хранимую рядом. Так появилась архитектура будущего решения.

[1] Подробнее: Сбер внедрил собственную графовую платформу. URL: https://www.sberbank.ru/ru/sberpress/all/article?newsID=6dc5a4d9-fffc-4d24-9de2-e038598a80c8&blockID=1303&regionID=77&lang=ru&type=NEWS (дата обращения: 06.05.2024).

Для того, чтобы возможности технологических разработок Сбера можно было применять для различных бизнес-задач, в том числе связанных с безопасностью данных, они были объединены в единую технологическую платформу, состоящую из семи сервисов:

Amber lab – сервис для разработки новых графовых моделей.

Ruby calc – сервис выполнения массовых расчетов на больших графах для инференса уже разработанных графовых моделей.

Jade API – сервис доступа к данным графа «на лету» в режиме «вопрос – ответ».

Link Prediction – сервис поиска неявных связей для предсказания неизвестных фактов о связанности узлов (например, об интересах, потребностях клиентов).

Crystal view – сервис визуального анализа больших графов.

Fastgraph – графовая база данных, используемая для хранения больших данных, в том числе персональных, представленных в виде графов.

Jasper GNN – регламентированный процесс обучения графовых нейронных сетей, ориентированный на конкретную бизнес-задачу выполняемую AI.

И это далеко не все сервисы для работы с графами. Но благодаря ним уже удалось достичь решения многих бизнес-задач, сделать графовые расчеты одним из инструментов обеспечения безопасности данных, противодействия фроду и кибермошенничеству. Графовые технологии помогают сегодня в решении задач по защите данных от утечек, в том числе персональных данных клиентов Сбера.

Справка

На «языке графов» вершины – это, например, здания, офисы, компьютеры, файлы, клиенты, сотрудники, а ребра между вершинами – передаваемые данные. Структурируя таким образом информацию, можно выявить аномалии в действиях компьютеров или людей, предотвратить кибератаки, фрод, определить уязвимые места в инфраструктуре.

Человек или машина?

Есть два классических подхода в работе с графами:

В первом случае граф анализирует человек, глядя на цепочки связей, выявляя дополнительную информацию за счет визуализации данных. Примерно так, как мы строим кратчайший маршрут, смотря на карту дорог или метро. В целевом виде продукт работает так: на входе специалист голосом или в виде команды подает задачу (например, «найди связи между такими-то заемщиками»), а на выходе – готовый граф с наглядной визуализацией, отображающей маршруты связанности конкретных заемщиков и вся информация о связанности. Этот способ, хотя и дает множество возможностей по обработке данных для бизнеса, однако не может быть массовым, так как ограничен скоростью работы человека по анализу результата.

Во втором случае уже машина проводит графовые расчеты и выдает результат, который интерпретируется требуемым образом. Часто такие расчеты объединены с работой искусственного интеллекта, который правильно применяет полученные данные. Это, как правило, более массовый сценарий.

Графовые эмбеддинги – эффективно и безопасно

В основу методов работы искусственного интеллекта с данными, представленными в виде графов, положен метод создания графовых эмбеддингов.

Графовый эмбеддинг – это свертка графов в вектор значений для выполнения задач моделирования. Чаще всего эмбеддинг содержит набор графовых метрик, отражающих геометрические и прочие свойства окружения исследуемого узла.

Интересный факт

Вы когда-нибудь задумывались, как социальная сеть предлагает вам «подходящих» друзей, о существовании которых вы могли даже не подозревать? Или как приложения на вашем телефоне предсказывают эффективный маршрут с учетом дорожных пробок еще до того, как вы выехали на дорогу? Ответ прост – в основе таких технологий лежат графы, представляющие данные как узлы и связи между ними. Есть графы и в рекомендательных системах. Использование графовых эмбеддингов в рекомендательных системах – это, в своем роде, пример эффективного использования графовых технологий в системе управления данными, когда информация о клиентах структурирована таким образом, что между ними можно выстроить органичные связи. Эти связи с помощью эмбеддингов для каждого клиента сворачивают в векторы признаков, позволяющих искусственному интеллекту в дальнейшем делать пользователям очень точные персонализированные предложения.

Существует два основных типа методов, используемых в рекомендательных системах[2]:

- методы совместной фильтрации;
- методы, основанные на моделях.

Метод совместной фильтрации может предсказать, например, рейтинг фильма, не зная атрибутов фильмов и пользователей. Чтобы решить эту задачу, метод факторизации графов объединяет метод, основанный на модели, с методом совместной фильтрации для повышения точности прогнозирования.

Методы факторизации графов широко используются во многих онлайновых рекомендательных системах. Факторизация графов – это модель на основе графов, которая может быть использована для представления предпочтений пользователей, а также для выстраивания связей между пользователями, предметами и атрибутами. Цель факторизации графов: извлечь скрытые характеристики из пользовательских оценок и рекомендаций, чтобы использовать их для прогнозирования предпочтений пользователей и, например, предложения релевантных потребностям продуктов, сервисов, услуг.

Факторизация графов выполняется путем разбиения исходного набора данных на более мелкие наборы или кластеры. Этот процесс может быть осуществлен с помощью баз данных графов, поскольку они разработаны для поддержки высокосвязанных структур данных и связей между точками данных.

Эмбеддинг дает такую же пользу для построения AI/ML в качестве входных фичей, как если бы на вход подавалась полноценные графовые данные, однако он прост, компактен и обезличен.

Такой вектор не хранит персональные данные (причем речь не только о прямых идентификаторах, но и о косвенных) и позволяет использовать эту свертку только для конкретной задачи без возможности последующего использования данных для других целей. Такие векторы можно использовать даже за пределами компании без риска утечки персональных данных.

Использование графовых вычислений при работе с AI/ML позволяют получать скрытые зависимости и выполнять предиктивный анализ информации для реализации AI-алгоритмов и получения ответов на запросы в режиме реального времени[3]. Поскольку автономная AI-модель направлена на решение конкретной односложной задачи, графовая аналитика работает в качестве подкрепляющего инструмента, позволяющего выстраивать комплексные зависимости данных. Например, для предсказания поведения групп людей в социальных сетях.

Заключение

Таким образом, для бизнеса крайне важно эффективное управление рисками, в том числе связанными с персональными данными, что сейчас для крупных корпораций является особым приоритетом. Уже сегодня графовая платформа Сбера[4] помогает бороться с утечками данных, предсказывать вероятные дефолты заемщиков, защищать карточные транзакции от фрода, делать точные таргетированные предложения клиентам и решать множество других задач за счет эффективной и безопасной обработки данных клиентов.

Графовая платформа Сбера претендует на то, чтобы стать бенчмарком на российском рынке и, возможно, сможет в будущем использоваться в качестве технологической базы для других компаний, что откроет им новые возможности по обработке больших данных и в то же время обеспечит поддержание стандартов управления данными.

[2] Wayne Sheng. Graph Database For Recommendation Systems. URL: https://www.nebula-graph.io/posts/use-cases-of-graph-databases-in-real-time-recommendation (дата обращения: 02.05.2024).

[3] Пилецкий И. И., Батура М. П., Шилин Л. Ю. Графовые технологии в интеллектуальной системе комплексного анализа данных интернет-источников. Доклады БГУИР. 2020. №5. URL: https://cyberleninka.ru/article/n/grafovye-tehnologii-v-intellektualnoy-sisteme-kompleksnogo-analiza-dannyh-internet-istochnikov (дата обращения: 03.05.2024).

[4] Графовой платформой Сбера пользуются 6500 специалистов. URL: https://sber.pro/digital/publication/grafovoj-platformoj-sbera-polzuyutsya-6500-speczialistov/ (дата обращения: 03.05.2024).

Олеся Сидоренко

Эксперт, Центр DPO, Сбер

Путешествие в страну Оз[1]

Опыт Австралии в управлении данными

Вступление

Сегодня, в эпоху активной цифровизации всех сфер общественной жизни, вопросы эффективного управления данными становятся все более актуальными.

Практически все страны мира признали важность данных и перехода к дата-центричной модели[2] государственного управления, в основе которой:

- развитие законодательства в области управления данными, нацеленного на регулирование передовых дата-ориентированных подходов[3];
- выстраивание экосистемы государственных и коммерческих данных, в которой возможно создание и развитие систем обмена данными между государством и бизнесом, а также их эффективное использование;
- совершенствование методов обезличивания данных;
- повышение ценности данных за счет обеспечения доверия к ним на основе точности и актуальности данных, возможности их безопасной обработки, доступности, этичности;
- совершенствование компетенций при работе с данными;
- повышение уровня цифровой грамотности населения в целом и представителей бизнес-сегмента.

При этом, несмотря на то, что большинство участников рынка осознают важность вопроса управления данными, на государственном уровне только лишь в отдельных странах имеется опыт законодательного урегулирования данного вопроса, что является проблемой в условиях стремительного увеличения количества данных, над которыми нужен контроль. В этой статье предлагаем поговорить о глобальных законодательных реформах Австралии в вопросе управления данными и узнать о причинах изменения законодательства, сути реформ и эффекте от их реализации, который ожидается в Австралии.

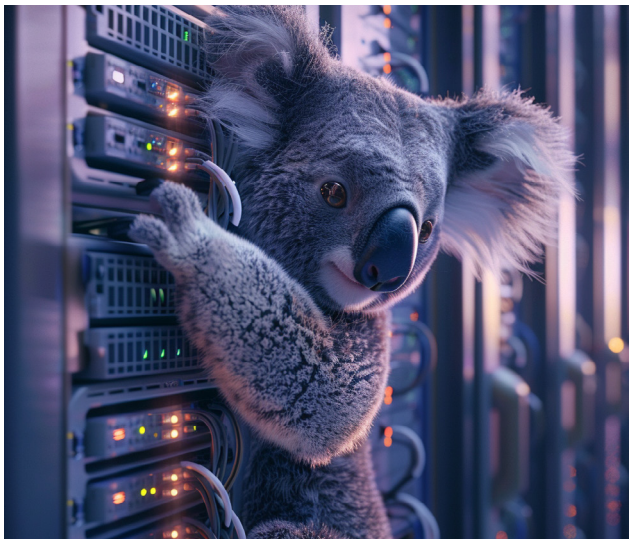
[1] Австралийцы называют себя «оззи», а свою страну Оз – по первым звукам от английского Australia.

[2] Дата-центризм – это архитектура построения информационных систем, где ключевым звеном являются данные.

[3] Дата-ориентированный подход предполагает принятие решений на основе точных и актуальных данных, которые предварительно были тщательно проанализированы.

Справка

В Российской Федерации (далее – РФ) создана Национальная система управления данными (далее – НСУД). Ее создание предусмотрено Федеральным проектом «Цифровое государственное управление» национальной программы «Цифровая экономика РФ»[4]. Основной целью создания и обеспечения функционирования НСУД является повышение эффективности создания, сбора и использования государственных данных как для предоставления государственных и муниципальных услуг, так и для осуществления государственных и муниципальных функций в электронной форме. Постановлением Правительства РФ от 14 мая 2021 г. №733 «Об утверждения Положения о федеральной государственной информационной системе «Единая информационная платформа национальной системы управления данными» и о внесении изменений в некоторые акты Правительства РФ» определен порядок использования НСУД, а также цели, задачи и состав участников единой информационной платформы. При этом Минэкономразвития России совместно с Минцифры России[5] ведется работа по разработке проекта федерального закона «О внесении изменений в Федеральный закон «Об информации, информационных технологиях и о защите информации» в части формирования национальной системы управления данными» с целью урегулирования отношений, связанных с передачей национальных данных, находящихся в распоряжении органов и организаций и полученных ими в связи с выполнением ими государственных и муниципальных функций.



[4] Национальная система управления данными. Министерство цифрового развития, связи и массовых коммуникаций РФ. URL: https://digital.gov.ru/ru/activity/directions/1061/ (дата обращения: 08.05.2024).

[5] «Ведомости» узнали о разработке принципов очистки платных для бизнеса «нацданных». URL: https://www.forbes.ru/tekhnologii/495413-vedomosti-uznali-o-razrabotke-principov-ocistki-platnyh-dla-biznesa-nacdannyh (дата обращения: 08.05.2024).

Глобальные австралийские реформы управления данными: как DATA поможет обеспечить доступность и прозрачность?

При изучении опыта зарубежных стран по вопросу управления данными интересным показался опыт Австралии, где была выработана собственная стратегия по управлению данными, на основе которой в 2022 году приняли отдельный Закон о доступности и прозрачности данных **(The Data Availability and Transparency Act 2022 (Cth)** (далее – DAT Act)[6], устанавливающий на законодательном уровне схему обмена данными в Австралии.

Впервые вопрос обмена данными в Австралии был поднят в мае 2017 года. В частности, австралийской Productivity Commission[7] был опубликован отчет «Data Availability and Use»[8], в котором подробно проанализировано текущее состояние государственных, частных, открытых, научных и иных видов данных, а также возможности, которые предоставляют данные для частных лиц, бизнеса, государства и общества в целом.

> Productivity Commission был сделан акцент на том, что данные позволяют создавать новые бизнес-модели, продукты и формировать новые идеи. Однако существующий подход к данными в Австралии не позволяет этого сделать, в связи с чем нуждается в реформе вопрос управления данными.

Это связано с тем, что улучшенный доступ к данным и их использование могут позволить создавать новые продукты и услуги, которые преобразуют повседневную жизнь современных граждан, повысят производительность, а также позволят принимать более эффективные решения в различных сферах жизни. При этом Productivity Commission было отмечено, что незначительных изменений будет недостаточно.

Поэтому ими был предложены рекомендации **по трансформации государственной политики в области данных**. По мнению вышеуказанного органа, в основе реформы должен быть переход от системы, основанной на неприятии риска, к системе, основанной на прозрачности и доверии к процессам обработки данных и рассматривающей данные как актив, а не угрозу[9].

[6] Data Availability and Transparency Act 2022. URL: https://www.legislation.gov.au/C2022A00011/latest/text (дата обращения: 08.05.2024).

[7] Главный аналитический и консультативный орган правительства Австралии по вопросам макроэкономической политики, регулирования и ряду других социальных и экологических проблем.

[8] Data Availability and Use. URL: https://www.pc.gov.au/inquiries/completed/data-access/report (дата обращения: 08.05.2024).

[9] Примечание: в данном случае не имеется в виду отсутствие потребности в обеспечении безопасности данных от внешних угроз. Позиция предполагает только исключение восприятия данных как угрозы, от которой нужно «уходить», например, за счет сокращения количества обрабатываемых данных.

Для трансформации государственного управления в области данных Productivity Commission **предложила**:

- принять Закон об обмене и разглашении данных;
- создать Национальный орган хранения данных для руководства и мониторинга новых механизмов доступа и использования, включая управление рисками;
- создать структуру обмена и разглашения данных, которая покажет всем хранителям данных, что взят курс на эффективное использование данных;
- образовать аккредитованных операторов данных с целью упрощения доступа к наборам данных, представляющих национальный интерес, а также к другим наборам данных, которые можно было бы связать и совместно использовать или обнародовать.

При этом Productivity Commission было отмечено: внедрение изменений потребует дополнительных затрат со стороны Правительства Австралии, однако они будут компенсированы открывающимися возможностями.

Подготовленный отчет был направлен в Правительство Австралии для рассмотрения. Ответ Правительства на запрос Productivity Commission не заставил себя долго ждать. Основные тезисы:

- данные – это национальный ресурс, который предоставляет гражданам, бизнесу и правительству колоссальные возможности для принятия исключительных решений и разработки инновационных продуктов и услуг;
- важно на государственном уровне пересмотреть подход в использовании данных, чтобы Австралия оставалась конкурентоспособной страной в современной глобальной экономике;
- необходимо инвестировать 65 млн австралийских долларов для реформирования австралийской системы данных и внедрения ряда мер по выполнению рекомендаций Productivity Commission[10].

В итоге было **принято решение** пойти дальше – **преобразовать систему данных** в Австралии и способы предоставления и использования данных.

В основу реализации реформы были положены **три ключевых тезиса**:

1. Прозрачность и контроль над собственными данными.
2. Национальный уполномоченный по обработке данных должен быть надежным надзирателем за системой публичных данных.
3. Новые законодательные и управленческие механизмы = эффективное использование данных в масштабах всей экономики.

Совершенствование австралийской системы данных – это новые возможности для стимулирования инноваций и расширения австралийской экономики. На такой оптимистичной ноте в 2018 году началась работа по реформированию системы данных в Австралии.

[10] The Australian Government's response to the Productivity Commission Data Availability and Use Inquiry. URL: https://dataavailability.pmc.gov.au/ministers-foreword.html (дата обращения: 08.05.2024).

DAT Act: новые правила обмена данными

В 2020 году впервые был **представлен проект** ранее упомянутого DAT Act, в котором была предложена схема контролируемого доступа к данным государственного сектора.

В пояснительной записке к законопроекту было отмечено, что **реформы необходимы** для реализации преимуществ большей доступности и использования данных, выявленных Productivity Commission, а также для поддержки экономических и исследовательских возможностей и эффективного предоставления услуг[11].

DAT Act был **принят в марте 2022 года и вступил в силу с 1 апреля**.

Он устанавливает новую **передовую схему обмена данными** (далее – DATA Scheme)[12], подкрепленную надежными гарантиями и последовательными эффективными процессами. DATA Scheme направлена на повышение доступности использования данных, что позволяет обеспечивать более эффективную государственную политику, а также поддерживать ведущие мировые исследования и разработки.

DAT Act предполагает создание DATA Scheme в государственном секторе, которая регламентирует вопрос обмена данными между органами правительства Австралии (Содружества), другими органами Содружества, государственными органами штата или территории Австралии, или университетами Австралии.

Среди ключевых аспектов регулирования:

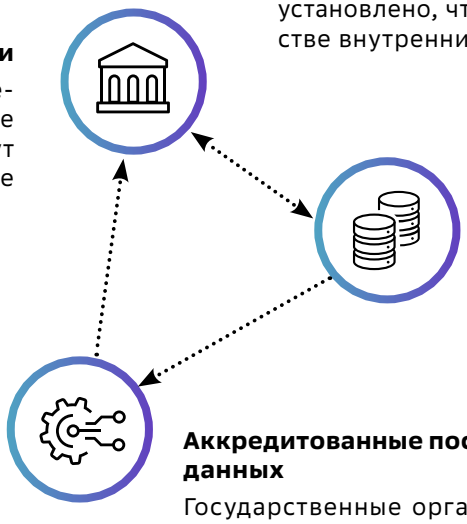
- участники информационного обмена должны быть ***аккредитованы***, прежде чем они смогут получать и использовать данные. После аккредитации они, в соответствии с DATA Scheme, получают статус аккредитованных пользователей;
- государственные органы Содружества, штатов и территорий, а также австралийские университеты, которые обеспечивают комплексную интеграцию данных, а также оказывают услуги по деидентификации, должны быть непосредственно ***аккредитованными поставщиками*** услуг по обработке данных;
- ***хранителями данных***, которые передают данные аккредитованным пользователям, ***являются государственные органы Австралии***. При этом хранители данных не отказываются от участия в DATA Scheme, а становятся ими автоматически и заявку на эту роль им подавать не нужно.

В соответствии с DATA Scheme, данные могут включать в себя широкий спектр тем: бытовые (например, погодные условия, урожайность хозяйства, движение транспорта), вопросы производства (например, грузоперевозки), наконец, что особенно важно для сферы персональных данных – личные данные участников информационного обмена и не только.

При этом, исходя из соображений национальной безопасности, некоторые организации были исключены из DATA Scheme, в частности, Австралийская федеральная полиция и Австралийская организация безопасности и разведки. В том числе установлено, что данные, хранящиеся в Министерстве внутренних дел, не могут быть переданы.

Аккредитованные пользователи

Государственные органы Содружества, штатов и территорий, а также австралийские университеты могут получать и использовать данные правительства Австралии



Хранители данных

Хранителями данных являются Государственные органы Австралии

Аккредитованные поставщики услуг обработки данных

Государственные органы Содружества, штатов и территорий, а также австралийские университеты, которые могут предоставлять специализированные услуги по обработке данных, такие как комплексная интеграция данных, деидентификация и/или услуги безопасного доступа к данным, для поддержки проектов по обмену данными

[11] New Australian Government data sharing legislation commences. URL: https://www.holdingredlich.com/new-australian-government-data-sharing-legislation-commences (дата обращения: 08.05.2024).

[12] The DATA Scheme. URL: https://www.datacommissioner.gov.au/the-data-scheme (дата обращения: 08.05.2024).

Согласно DAT Act **любой обмен данными в рамках схемы должен:**

1 Использоваться для одной или нескольких целей обмена данными

В соответствии с DATA Scheme обмен данными должен осуществляться в интересах общества и для одной из трех целей:

- предоставление государственных услуг;
- информирование о государственной политике и программах;
- исследования и разработки.

При этом важно отметить, что DAT Act устанавливает запрет на обмен данными в рамках схемы для целей, связанных с правоохранительными органами, или в целях национальной безопасности.

2 Соответствовать принципам обмена данными

Принципы обмена данными определяют «проект», «людей», «условия», «данные» и «результаты» в качестве соответствующих параметров для оценки запросов на обмен данными и управления соответствующими рисками.

В частности, при оценке запроса необходимо установить следующее:

1. Проект направлен на реализацию общественных интересов.

2. Доступ к данным предоставляется только лицам, которые обладают качествами, квалификацией, связями или опытом, необходимыми для доступа.

3. Обмен данными, их сбор и использование осуществляются в надлежащим образом контролируемой среде:

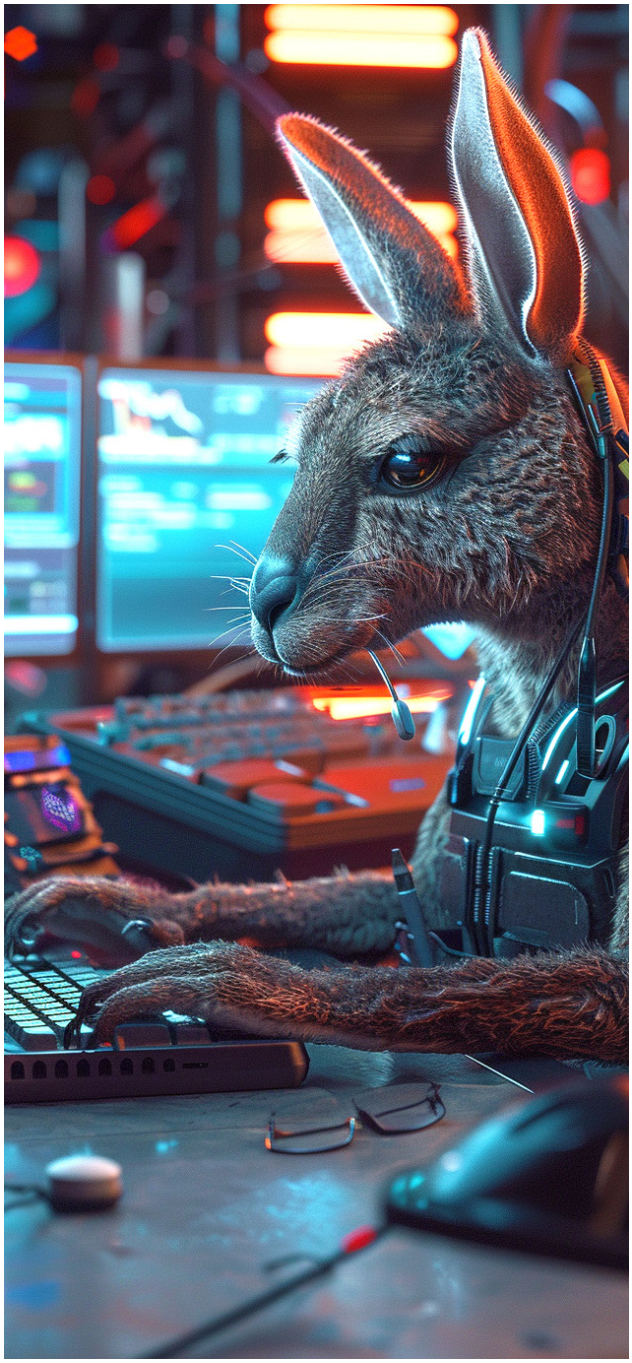
- средства, с помощью которых передаются, собираются и используются данные, являются надлежащими, с учетом типа и чувствительности данных, для контроля рисков несанкционированного использования;
- при обмене, сборе и использовании данных применяются разумные стандарты безопасности.

4. Передаются, собираются и используются только те данные, которые необходимы для достижения применимой цели или задач обмена данными.

5. Единственным результатом проекта является получение конечного результата и/или выходных данных.

3 Осуществляться в рамках соглашения об обмене данными

Соглашение об обмене данными касается обмена данными государственного сектора. Сторонами такого соглашения являются хранитель данных государственного сектора и аккредитованный пользователь.



Data ACT VS Privacy Act: соотношение регулирования в области управления данными и регулирования в области защиты частной жизни, персональных данных

Важно, что DAT Act не отменяет Закон о конфиденциальности 1988 года (Privacy Act 1988 – закон Австралии, регулирующий вопросы неприкосновенности частной жизни, в том числе определяющий требований к обработке персональных данных)[13]. Поэтому обмен, сбор и использование данных в рамках DATA Scheme должны соответствовать Закону о конфиденциальности 1988 года. В данном случае приоритет отдается более специальному регулированию, как и в России.

Справка

Интересно отметить, что Закон о конфиденциальности Австралии 1988 года сейчас также подвержен масштабному пересмотру в рамках реформы австралийского законодательства о неприкосновенности частной жизни. Австралия стремится урегулировать раскрытие персональных данных таким образом, чтобы сбалансировать частную жизнь своих граждан и конкурирующие общественные интересы (например, определить границы допустимости опубликования персональных данных в открытых источниках)[14].

Как уже ранее было отмечено, передаваемые данные в рамках DATA Scheme могут включать в том числе и личные данные. В этой связи DAT Act не только должен соответствовать специальному закону, регулирующему вопросы конфиденциальности данных, включая персональные данные, но еще и в нем самом установлены дополнительные меры по защите конфиденциальности, которые бы позволили свести к минимуму обмен личной информацией (то есть исключить случаи сбора и передачи избыточных данных).

В частности, DAT Act закреплено следующее:

- данные, включающие биометрические персональные данные, не должны передаваться, если физическое лицо, к которому относятся биометрические персональные данные, прямо не согласится на это;
- если передаются данные, содержащие личную информацию (персональные данные), то в соглашении об обмене данными, должен быть установлен запрет любой аккредитованной организации, с которой или через которую они передаются, хранить или предоставлять доступ к данным за пределами Австралии (территориальное ограничение);
- если передаются данные, которые были деидентифицированы, то соглашение об обмене данными должно запрещать получателю данных повторно идентифицировать данные.

Дополнительно DAT Act предусмотрены меры по защите конфиденциальности для конкретных целей.

Например, предоставляя государственные услуги, допустимо использовать данные, включающие личную информацию, но при наличии **одного или нескольких условий**, в частности:

- физическое лицо само предоставило согласие на передачу своей личной информации (то есть должен быть явный факт волеизъявления);
- передача осуществляется ввиду чрезвычайной ситуации и/или стихийного бедствия;
- предоставляемая услуга указана непосредственно в соглашении об обмене;
- предоставляется только минимальный объем личной информации, необходимый для надлежащего предоставления услуги.

Другой пример: если обмен данными, включающими персональные данные, осуществляется в целях информирования о государственной политике и программах или для проведения исследований и реализации разработок, то:

Либо должны быть совокупно выполнены условия № 1 (перечислены ниже):

- должно быть согласие физического лица на передачу его личной информации;
- передается только минимальный объем личной информации, необходимый для продолжения проекта.

Либо должны быть совокупно выполнены условия № 2 (перечислены ниже):

- проект не может продолжаться без личной информации;
- общественные интересы, которым служит проект, оправдывают передачу личной информации о физических лицах без их согласия;
- предоставляется только минимальный объем личной информации, необходимый для продолжения проекта;
- существуют разрешенные обстоятельства (например, когда запрашивать согласие физического лица неразумно или практически невозможно, или же когда передача осуществляется ввиду чрезвычайной ситуации и при стихийных бедствиях).

[13] Data Protection Act. 1988. URL: https://www.legislation.gov.au/C2004A03712/latest/text (дата обращения: 08.05.2024).

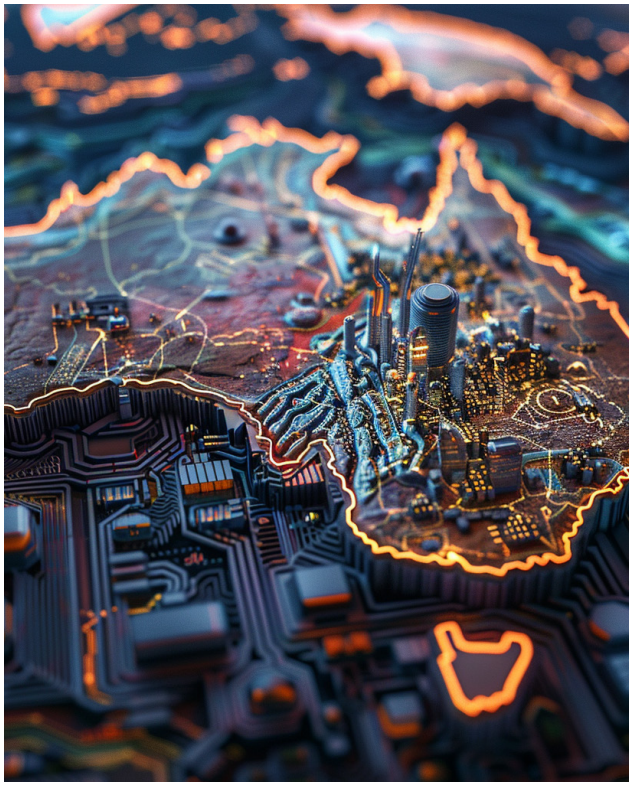
[14] Normann Witzleb. Responding to Global Trends? Privacy Law Reform in Australia. URL: https://www.degruyter.com/document/doi/10.1515/9783111010601-009/html (дата обращения: 08.05.2024).

Справка

Если хранитель передаваемых данных пришел к выводу от том, что согласие физического лица запросить неразумно или практически невозможно, то в соглашении об обмене данными должно быть указано, что личная информация передается без согласия отдельных лиц, поскольку запрашивать их согласие неразумно или практически невозможно, но при этом в соглашении об обмене должны быть объяснены причины, по которым хранитель данных пришел к такому выводу.

При этом в DAT Act есть примечание: нет ничего необоснованного или невыполнимого в том, чтобы запрашивать согласие отдельного лица только потому, что необходимо запросить согласие очень большого числа лиц, из чего следует, что обоснование причин, при которых неразумно или практически невозможно получить согласие физического лица на передачу, должно быть весомым.

Таким образом, при осуществлении обмена данных необходимо учитывать не только требования DAT Act, но и Закона о конфиденциальности 1988 года, чтобы сбор, использование и раскрытие личной информации соответствовало австралийскими принципами конфиденциальности.



Национальный уполномоченный по обработке данных

Отдельно стоит обратить внимание на то, что DAT Act предусматривает создание Национального уполномоченного по данным (далее – Уполномоченный) и соответственно Офиса Уполномоченного (далее – ONDC)[15]. Уполномоченный осуществляет надзор за DATA Scheme, установленной DAT Act, в том числе консультирует по ней и обеспечивает ее соблюдение в соответствии с Законом о конфиденциальности 1988 года и применимыми гарантиями безопасности.

В функции ONDC входит:

- аккредитация участников информационного взаимодействия;
- рассмотрение жалоб;
- оценка, расследование, принятие мер по обеспечению соблюдения DAT Act;
- передача дел в другой уполномоченный орган;
- ведение публичных реестров:
  - соглашений об обмене данными;
  - аккредитованных пользователей;
  - аккредитованных поставщиков услуг по обработке данных;
- подготовка ежегодного отчета о работе DATA Scheme, а также о своей деятельности и деятельности Национального консультативного совета по данным.

За первый полный год работы Уполномоченного, после вступления в силу DAT Act 1 апреля 2022 года, на основе передовой практики обмена данными, собранными правительственными учреждениями Австралии, были определены три ключевых приоритета на 2023-24 годы, а именно:

1. поощрять и поддерживать внедрение DATA Scheme;

2. аккредитовать и контролировать участников DATA Scheme;

3. обучать и направлять на основе имеющейся передовой практики обработки данных и обмена ими[16].

Интересно отметить, что в первый год своей работы ONDC был сосредоточен именно на том, чтобы **первоначально повысить осведомленность субъектов**, имеющих право на участие в DATA Scheme, а не на том, чтобы начать ее активно применять с первого дня.

[15] Office of the National Data Commissioner. URL: https://www.datacommissioner.gov.au/about-us/about-the-ondc (дата обращения: 08.05.2024).

[16] Annual Priorities 2023-24. URL: https://www.datacommissioner.gov.au/about-us/about-the-ondc/our-priorities (дата обращения: 08.05.2024).

В результате проведенной работы на конец июня 2023 года для участия в DATA Scheme было аккредитовано десять организаций, четыре из которых аккредитованы как пользователи, то есть им предоставлено право запрашивать и использовать данные, собранные правительством Австралии, а шесть из которых были аккредитованы как поставщики услуг, и соответственно могут теперь предоставлять такие услуги как деидентификация и интеграция данных, а также предоставлять услуги по обеспечению безопасного доступа к данным[17].

> Помимо прочего, после вступления в силу DAT Act одним из приоритетов для ONDC стало создание институциональных механизмов и инструментов поддержки безопасного, прозрачного и последовательного обмена данными.

В связи с чем ONDC были подготовлены два руководящих документа: Кодекс доступности и прозрачности данных 2022[18] и Кодекс доступности и прозрачности данных (меры национальной безопасности) 2022[19], в которых детализированы принципы обмена данными, основанные на передовой практике, включая проверку на соответствие общественным интересам, более подробно описаны подходы к согласиям, конфиденциальности и этике, а также изложен подход к управлению рисками национальной безопасности.

Dataplace

Помимо прочего, ONDC была создана цифровая платформа – Dataplace[20]. Она предназначена для правительственных учреждений (Содружества, штата, территории и местного самоуправления), а также австралийских университетов, исследователей и организаций частного сектора, которые хотят получить доступ к данным правительства Австралии.

Платформа необходима для того, чтобы объединить всех, кто хочет получить доступ к данным правительства Австралии (например, исследователей и тех, кто занимается государственной политикой и предоставлением государственных услуг), с агентствами Содружества, которые являются хранителями данных.

Dataplace можно использовать для:

- подачи заявок на аккредитацию в качестве пользователя данных или в качестве поставщика услуг передачи данных в соответствии с DATA Sheme;
- запроса данных правительства Австралии;
- разработки соглашений об обмене данными;
- контроля своих действий по обмену данными.

Таким образом, данная цифровая платформа фактически упрощает процесс взаимодействия участников DATA Scheme друг с другом при запросах данных, особенно там, где у набора данных может быть несколько хранителей.

Data Inventories Pilot Program

ONDC занимает проактивную позицию в вопросе совершенствования процесса обмена данными в Австралии. Для облегчения обнаружения, использования и повторного использования государственных данных ONDC было принято решение о разработке Каталога данных правительства Австралии, который направлен на повышение прозрачности государственных хранилищ данных, сокращение дублирования данных и обеспечение более широкого повторное использования данных и обмена ими[21].

Каталог будет основан на перечнях данных агентств, разработанных в рамках Data Inventories Pilot Program (далее – DIPP), а также на открытых и других источниках данных, которые помогут пользователям находить данные правительства Австралии. В рамках DIPP ONDC сотрудничает с правительственными учреждениями Австралии, помогая им находить свои данные и разрабатывать стандартизированный список хранящихся у них информационных ресурсов, известный как инвентаризация данных. DIPP реализуется в рамках Стратегии управления данными и цифровыми технологиями на период до 2030 года (Data and Digital Government Strategy – DDGS)[22].

Новая стратегия направлена на цифровое преобразование, ориентированное на предоставление услуг для всех людей и предприятий, предоставление простых и бесперебойных услуг, а также создание информационной и цифровой основы для поддержки правительства «на будущее», которому «доверяют» и которое «защищают».

[17] Annual Report 2022-23. URL: https://www.datacommissioner.gov.au/sites/default/files/2023-10/ONDC%20Annual%20report.pdf (дата обращения: 08.05.2024).

[18] Data Availability and Transparency Code 2022. URL: https://www.legislation.gov.au/F2022L01719/asmade/text (дата обращения: 08.05.2024).

[19] Data Availability and Transparency (National Security Measures) Code 2022. URL: https://www.legislation.gov.au/F2022L01722/latest/text (дата обращения: 08.05.2024).

[20] Dataplace. URL: https://www.datacommissioner.gov.au/use-dataplace (дата обращения: 08.05.2024).

[21] Australian Government Data Catalogue. URL: https://www.dataanddigital.gov.au/plan/roadmap/delivering-for-all-people-and-business/australian-government-data-catalogue (дата обращения: 08.05.2024).

[22] Transforming Australia's Digital Governance. URL: https://opengovasia.com/2023/12/22/transforming-australias-digital-governance/ (дата обращения: 08.05.2024).

Заключение

Законодательство Австралии в области управления данными никогда не было объектом пристального внимания у исследователей, поэтому в этой статье была предпринята попытка исследовать ключевые аспекты данного вопроса.

Основные выводы, которые можно сделать:

1. Австралия участвует в мировых тенденциях в области управления данными, но все же выбирает свой собственный путь. Так, Австралия не придерживается ни относительно строгого подхода Европейского союза, ни более разрешительного подхода США. От Российской Федерации подходы Австралии во многих вопросах также отличаются: хотя общие регуляторные тенденции также есть (например, в части создания платформ обмена информацией).
2. Австралийский DAT Act является примером того, что Австралия пошла по собственному пути создания современной системы обмена данными с акцентом на эффективное взаимодействие за счет доступности данных. Например, предоставление австралийским университетам права на пользование данными в лице аккредитованных пользователей является важным шагом на пути к более широкому использованию данных, наряду с надлежащей технологической поддержкой, этическими принципами и надзором. Если исторически государственные органы ограничивали доступ исследователей к данным, теперь предоставление доступа к правительственным данным может создать предпосылки для повышения качества исследований в Австралии.
3. При этом гармонизация новых регуляторных веяний, в основе которых доступность данных, создается за счет подчинения нового закона законодательству о защите частной жизни и персональных данных, что немаловажно для обеспечения конфиденциальности личных данных граждан Австралии.

Таким образом, стремительно ускоряющийся технологический прогресс и признание ценности данных лежат в основе реализуемых в разных странах, в том числе в Австралии, дата-центричных подходов управления данными.

Между тем Россия также не отстает от мировых трендов в области управления данными. Так, например, Стратегией развития информационного общества в РФ на 2017-2030 годы, утвержденной Указом Президента РФ от 9 мая 2017 г., отмечено, что главным способом обеспечения эффективности цифровой экономики[23] является внедрение технологии обработки данных, а конкурентным преимуществом на мировом рынке обладают государства, отрасли экономики которых основываются на технологиях анализа больших объемов данных[24]. В связи с чем Россия взяла курс на активное развитие цифровой экономики и внедрение ряда проектов в данной области. Так, например, в рамках национальной программы «Цифровая экономика РФ» реализовывается федеральный проект «Цифровое государственное управление», в соответствии с которым создана ранее упомянутая НСУД. Ее создание является важным шагом на пути решения существующих проблем в области управления данными, поскольку позволяет в первую очередь повысить эффективность использования данных и сократить финансовые и временные затраты на их обработку.

Кроме того, с учетом новых современных вызовов и стремительной цифровизации общества Президентом РФ было поручено Правительству РФ разработать новый национальный проект «Экономика данных и цифровая трансформация государства» на 2025-2030 годы[25]. Предполагается, что он придет на смену национальной программы «Цифровая экономика», а его задачи будут значительно шире. Новый национальный проект будет устремлен на более широкое использование данных и дальнейшую цифровую трансформацию государственных структур. Как отметил глава Минцифры России Максут Шадаев: «Приоритетов и задач стало больше, многие показатели стали гораздо более амбициозными»[26]. С учетом важности реализации инициатив в области управления данными на основе принципов конфиденциальности, а также требований, предъявляемых к безопасности данных для минимизации рисков реализации современных угроз в отношении чувствительной информации, среди которой персональные данные, в рамках нового национального проекта вопрос по обеспечению информационной безопасности выделен отдельно как одно из приоритетных направлений в области цифровой трансформации общества.

---

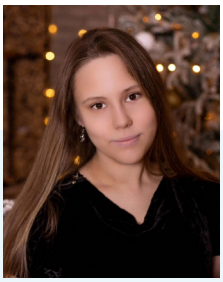
[23] Цифровая экономика – это хозяйственная деятельность, в которой ключевым фактором производства являются данные в цифровом виде, обработка больших объемов и использование результатов анализа которых по сравнению с традиционными формами хозяйствования позволяют существенно повысить эффективность различных видов производства, технологий, оборудования, хранения, продажи, доставки товаров и услуг (Источник: Указ Президента РФ № 203).

[24] Указ Президента РФ от 9 мая 2017 г. № 203 «О Стратегии развития информационного общества в Российской Федерации на 2017-2030 годы». URL: https://base.garant.ru/71670570/ (дата обращения: 08.05.2024).

[25] Перечень поручений по реализации Послания Президента Федеральному Собранию от 30. 03.2024 № Пр-616. URL: http://www.kremlin.ru/acts/assignments/orders/73759 (дата обращения: 08.05.2024).

[26] Максут Шадаев: «Экономика данных» идет на смену «Цифровой экономике». URL: https://ict-online.ru/news/Maksut-Shadayev-Ekonomika-dannykh-idet-na-smenu-Tsifrovoi-ekonomike-289850 (дата обращения: 08.05.2024).

Тренды и вызовы в области обработки данных и управления Big Data



Полина Сурьянинова

Редактор,
Аналитик, Центр DPO, Сбер

В современных технологических реалиях управление стремительно увеличивающимся количеством данных стало важнейшим компонентом успешности бизнеса. Прогнозируется бум приобретения данных для обучения AI и гиперавтоматизации[1]. Предлагаем вам актуальные цифры и статистику, чтобы вы сами могли в этом убедиться.

Глобальное распространение интернета

По данным Минцифры России[2], в первом квартале 2023 года аудитория интернета в России увеличилась до **101,4 млн** человек,

а это **83 %** населения страны.

Даже просто при запуске работы браузера – данные о нас уже попадают в сеть.

Как собираются данные о вас?

70 % данных создают сами пользователи.

Каждое наше действие в сети оставляет цифровой след[3] – будь то размещение фото в социальных сетях, оплата онлайн-покупок, добавление музыки в избранное, время, проведенное на сайте и многое другое.

Для желающих научиться защищать свои данные в интернете или для тех, кто обучает этому население, наш новый онлайн-курс «Вселенная персональных данных: защита в интернете». Курс доступен на платформе финансовой грамотности Сбера «СберСова» по ссылке.

[1] Источник: РБК Тренды. URL: https://trends.rbc.ru/trends/industry/65671a3f9a79472c43a1ec1c?from=copy (дата обращения: 13.05.2024).

[2] Источник: gov.ru. URL: https://digital.gov.ru/uploaded/files/internet-v-rossii-v-2022-2023-godah.pdf (дата обращения: 13.05.2024).

[3] Источник: Demandsage. URL: https://www.demandsage.com/big-data-statistics/ (дата обращения: 13.05.2024).

Количество обрабатываемых данных

Современные компании собирают данные с беспрецедентной скоростью. Каждый день создается примерно **2,5 квинтиллиона** байт данных[4].

Прогнозируется, что к 2025 году число возрастет до **180 зеттабайт.**

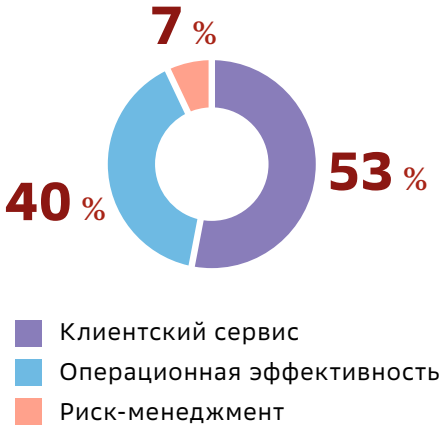
Среди такого объема данных также – персональные данные. Только представьте, что было бы в отсутствие управления такими объемами как на государственном уровне, так и на уровне бизнеса.

Сферы применения больших данных

Большие данные – это огромные массивы данных, которые настолько сложны и обширны, что не могут быть интерпретированы человеком или традиционными системами управления данными. Они формируются, в том числе, на основе цифровых следов, оставляемых пользователями в интернете. При правильном анализе с помощью современных инструментов эти массивы предоставляют бизнесу информацию, необходимую для принятия обоснованных решений.

Согласно исследованию IBM[5], большие данные используются в трех основных сферах:

- Клиентский сервис – это действия, направленные на поддержку клиентов на каждом этапе взаимодействия.
- Операционная эффективность – это оптимальное соотношение между ресурсами и результатом.
- Риск-менеджмент – это система управления, ориентированная на разработку и реализацию экономически обоснованных для бизнеса решений по устранению имеющихся рисков (например, правовых, регуляторных, репутационных и иных).



Каждое из этих действий сопряжено с анализом персональных данных пользователей – именно они помогают бизнесу персонифицировать свои продукты, сервисы, услуги для клиентов, а также принимать эффективные решения, направленные на увеличение прибыли и оптимизацию оборота.

[4] Источник: CloudTweaks. URL: https://cloudtweaks.com/2015/03/how-much-data-is-produced-every-day/ (дата обращения: 13.05.2024).

[5] Источник: IBM Institute for Business Value. URL: https://www.ibm.com/thought-leadership/institute-business-value/en-us (дата обращения: 13.05.2024).

Облачная миграция данных как новый тренд в управлении

По данным Gartner, к 2025 году

>85 % компаний

перейдут на облачные технологии и не смогут в полной мере реализовать свои цифровые стратегии без использования облачных архитектур и технологий[6].

С точки зрения управления данными тренд на «облака» – это определенный вызов, который нужно учитывать.

Постоянно увеличивающееся число операторов персональных данных

В Реестре операторов персональных данных Роскомнадзора содержатся сведения о более, чем **950 674** операторах[7].

К моменту выхода данного выпуска их будет еще больше. При этом не стоит забывать, что некоторые операторы могли еще не подать уведомление в Роскомнадзор, на основании которого производится включение в Реестр.

Представьте, какое количество персональных данных в различных целях обрабатывается этими операторами. В отсутствие эффективного управления данными, наряду с обеспечением их безопасности, повышается риск утечек персональных данных, их неправомерного попадания в открытый доступ, а соответственно риск привлечения к ответственности за нарушение прав субъектов.

Утечки персональных данных как одна из ключевых проблем современности

Утечки персональных данных – это в своем роде триггер для совершенствования систем управления данными наряду с системами безопасности.

С каждым годом число утечек персональных данных увеличивается.

По данным Роскомнадзора, количество утечек в России за первое полугодие 2023 года в 4 раза превысило показатели 2022 года[8] и составило **1,12 млрд** записей,[9]

а в 2024 году уже произошла утечка более **500 млн** записей о россиянах.

[6] Источник: Gartner. URL: https://www.gartner.com/en/newsroom/press-releases/2021-11-10-gartner-says-cloud-will-be-the-centerpiece-of-new-digital-experiences (дата обращения: 13.05.2024).

[7] Данные по состоянию на 14.06.2024. Актуальные данные: https://pd.rkn.gov.ru/operators-registry/operators-list/ (дата обращения: 14.06.2024).

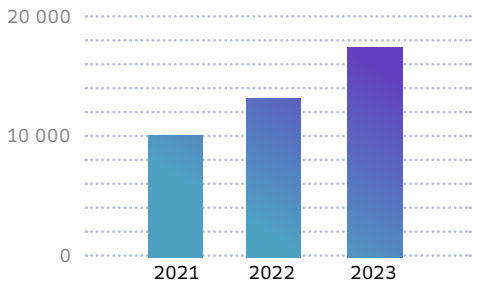
[8] Источник: Коммерсантъ. URL: https://www.kommersant.ru/doc/6443525 (дата обращения: 17.05.2024).

[9] Источник: РБК. URL: https://www.rbc.ru/society/11/03/2024/65ec41e89a7947dc41bd43f9 (дата обращения: 17.05.2024).

Как отметил Милош Вагнер: «В этом году одним этим случаем покрыли практически весь предыдущий год»[10]. Отсутствие эффективного управления данными, наряду с обеспечением защиты данных – одна из причин неправомерной обработки и утечек.

Утечки являются одной из причин, по которым увеличивается количество судебных дел, связанных с персональными данными:

Статистика судебных споров



Если в 2021 году в России было зафиксировано **10 256** судебных споров,

то в 2022 году их было уже **13 348**, на **30 %** больше,

а в 2023 – уже **17 497**, на **31 %** больше, чем годом ранее.

Итого: за два года рост составил **71 %.**[11]

На этом фоне более **2 тысяч** россиян обратились за бесплатной юридической помощью в Центр правовой помощи гражданам в цифровой среде в I квартале 2024 года,

что в **3 раза** превышает число заявителей за аналогичный период 2023 года[12].

Минимизация риска утечек персональных данных – общая задача для оператора и субъекта персональных данных. Оператору важно грамотно выстроить процессы управления данными и реализовать необходимые меры безопасности, чтобы исключить неправомерный доступ к информации и, как следствие, утечку данных. Субъектам же нужно быть внимательными при передаче своих данных: исключать избыточность передаваемых данных, быть начеку при получении писем от незнакомцев и не только.

[10] Источник: ТАСС. URL: https://tass.ru/obschestvo/20065237 (дата обращения: 13.05.2024).

[11] Источник: COMNEWS. URL: https://clck.ru/3AjL7a (дата обращения: 13.05.2024).

[12] Источник: ТАСС. URL: https://4people.grfc.ru/news/tass-chislo-obrativshihsya-v-centr-pravovoy-pomoschi-grazhdanam-v-cifrovoy-srede-vyroslo-v-tri-raza/ (дата обращения: 13.05.2024).

Обучение сотрудников – один из приоритетов в вопросах управления данными

Важно понимать: ряд утечек происходит по вине сотрудников организаций, допустивших ошибки в ходе реализации процесса управления данными.

По данным InfoWatch, в финансовом секторе в 2023 году по вине сотрудников произошло **~ 6 %** утечек[13].

В свою очередь, по данным Data maturity index,

64 % организаций отмечают, что их сотрудникам не хватает грамотности в работе с данными, поэтому обучение сотрудников должно быть в центре внимания компаний в ближайшие несколько лет[14].

Улучшение навыков управления данными позволит компаниям выстроить эффективную, экономичную и безопасную организацию процессов сбора, хранения и использования данных для принятия наиболее выгодных решений и обеспечения защиты прав и интересов субъектов персональных.

Полезную информацию по вопросу выстраивания эффективной системы организации обработки и защиты персональных данных, в основе которой управление данными, вы можете найти здесь в блоке «Экспертам».

Повышение цифровой грамотности населения как важный аспект предупреждения утечек

Проведенные исследования показали, что

66 % населения России отмечает нехватку знаний о цифровой грамотности и безопасности в сети[15].

Развитие культуры приватности необходимо, поскольку

1/3 населения сталкивается с интернет-мошенничеством, что приводит не только к финансовым потерям, но и к утечкам персональных данных.

Для тех, кто хочет всегда оставаться на страже личных данных или обучает этому население – новый проект Команды DPO Сбера «DPO на связи». С материалами можно ознакомиться на платформе финансовой грамотности Сбера «СберСова» по ссылке.

[13] Источник: InfoWatch. URL: https://www.infowatch.ru/sites/default/files/analytics/files/finansoviy-sektor-utechki-konfidentsialnoy-informatsii-za-tri-goda-mir-rossiya.pdf (дата обращения: 13.05.2024).

[14] Источник: Data maturity index. URL: https://carruthersandjackson.com/data-maturity-index/ (дата обращения: 13.05.2024).

[15] Источник: gov.ru. URL: https://digital.gov.ru/uploaded/files/internet-v-rossii-v-2022-2023-godah.pdf (дата обращения: 17.05.2024).

Олеся Сидоренко

Ответственный за дайджест, Эксперт, Центр DPO, Сбер

Privacy-Дайджест

Аналитика и обзор ключевых новостей в области персональных данных за второй квартал 2024 года

ВСТУПИЛИ В СИЛУ ВО ВТОРОМ КВАРТАЛЕ 2024 ГОДА:

01 Близкие родственники лица, пропавшего без вести, теперь подлежат обязательной государственной геномной регистрации

Федеральный закон от 14.02.2024 №16-ФЗ «О внесении изменений в Федеральный закон «О государственной геномной регистрации в Российской Федерации», Постановление Правительства РФ от 11.04.2024 №455 «О внесении изменений в постановление Правительства РФ от 24 июня 2023 г. №1027»

Суть изменений:

С 15 мая 2024 года родители, дети, полнородные братья или сестры лица, пропавшего без вести, подлежат обязательной государственной геномной регистрации.

Справка

Государственная геномная регистрация – деятельность по получению, учету, хранению, использованию, передаче и уничтожению биологического материала, а также обработке **геномной информации**.

Геномная информация – биометрические персональные данные, включающие кодированную информацию об определенных фрагментах ДНК физического лица или неопознанного умершего.

Как это будет происходить:

- *на основании обращения органа предварительного следствия или органа дознания родители, дети, полнородные братья и сестры[1] пропавшего без вести должны сдать свой биологический материал, содержащий геномную информацию;*
- *цель обработки такой информации будет заключаться в проведении государственной геномной регистрации;*
- *хранение геномной информации будет осуществляться до установления места нахождения пропавшего без вести, но не более 70 лет;*
- *уничтожение геномной информации осуществляется по истечении срока хранения или в случае установления места нахождения лица, пропавшего без вести, или в случае идентификации умершего.*

Ссылка

Ссылка

[1] Получение биологического материала от лиц, признанных недееспособными, или несовершеннолетних лиц, осуществляется в присутствии законных представителей таких лиц.

ОПУБЛИКОВАНЫ ВО ВТОРОМ КВАРТАЛЕ 2024 ГОДА:

01 Во исполнение требований 572-ФЗ принято новое Постановление Правительства РФ

Постановление Правительства РФ от 01.04.2024 № 408 «О видах биометрических персональных данных, на которые распространяется действие Федерального закона № 572-ФЗ[2]» (вступает в силу с 01.09.2024 и действует до 01.09.2027)

Суть изменений:

Во исполнение требований ч. 9 ст. 26 Закона № 572-ФЗ, которой предусмотрено, что с 1 сентября 2024 года положения Закона №572-ФЗ распространяются только на виды биометрических персональных данных, утвержденных Правительством РФ, принято Постановление Правительства РФ № 408. Перечень видов биометрических персональных данных, утвержденных ранее Законом № 572- ФЗ, остался прежним.

Ссылка

02 Согласия на размещение и обработку персональных данных в ЕСИА и биометрических персональных данных в ЕБС, а также на передачу векторов ЕБС теперь собираются по новым формам

Распоряжение Правительства РФ от 09.04.2024 №856-р «О внесении изменений в Распоряжение Правительства РФ от 30.06.2018 №1322-р» (вступает в силу с 01.01.2025)

Суть изменений:

С 1 января 2025 года на размещение и обработку персональных данных в ЕСИА и биометрических персональных данных в ЕБС, в том числе на передачу векторов ЕБС, необходимо будет использовать новые формы согласия. Если сейчас предусмотрена только одна форма, то будет две новых, которые различаются в зависимости от способа, которым предоставляется согласие:

- *в электронном виде;*
- *на бумажном носителе.*

В новых формах:

- *дополнительно предусмотрено поле для указания информации о несовершеннолетних лицах, чьи персональные данные размещаются в ЕСИА и ЕБС – заполняется в случае предоставления согласия на размещение и обработку персональных данных несовершеннолетнего лица его законным представителем;*
- *детализированы случаи, когда действие согласия прекращается:*
  - *действие согласия на обработку персональных данных оператором ЕСИА прекращается в день удаления оператором учетной записи субъекта в ЕСИА на основании заявления субъекта об удалении учетной записи (неприменимо в случае регистрации в ЕСИА и при самостоятельном размещении субъектом биометрических персональных данных в ЕБС с использованием мобильного приложения);*

[2] Федеральный закон от 29.12.2022 № 572-ФЗ «Об осуществлении идентификации и (или) аутентификации физических лиц с использованием биометрических персональных данных, о внесении изменений в отдельные законодательные акты Российской Федерации и признании утратившими силу отдельных положений законодательных актов Российской Федерации».

- *действие согласия на обработку биометрических персональных данных и персональных данных оператором ЕБС прекращается либо в день, который следует за днем истечения срока использования биометрических персональных данных, размещенных в ЕБС; либо в срок, не превышающий 30 дней с даты поступления отзыва[3]; либо в день предоставления оператору ЕБС требования об уничтожении или удалении биометрических персональных данных с использованием Госуслуг;*
- *действие согласия на обработку биометрических персональных данных и биометрических персональных данных, указанных в пп. «в» п. 3 согласия в бумажной форме и в п. 3 согласия в форме электронного документа, прекращается по достижении цели их обработки (размещение биометрических персональных данных в ЕБС и персональных данных в ЕСИА).*

Ссылка

[3] Если иное не предусмотрено договором, стороной которого, выгодоприобретателем или поручителем по которому является субъект персональных данных, иным соглашением между оператором и субъектом персональных данных либо если оператор не вправе осуществлять обработку персональных данных без согласия субъекта персональных данных на основаниях, предусмотренных Законом № 152-ФЗ или другими федеральными законами.

Рекомендации
от редколлегии[1]

Почему управление данными играет такую важную роль в текущих реалиях? Оно помогает выстроить процессы, в которых обрабатываются данные, эффективно, экономично и безопасно. Более того, оно позволяет оптимизировать принятие решений на основании управляемых данных для удовлетворения интересов как клиентов, так и бизнеса. Но что может произойти, если кто-то вмешается в установленный процесс управления данными, и к каким последствиям это может привести?

ЧТО ПОСМОТРЕТЬ?

Безос. Человек, создавший Amazon

2023

История начинается с того, что мы знакомимся с успешным бизнесменом Джеффом Безосом. Он завоевал репутацию одного из самых молодых и успешных людей на Уолл-стрит, однако не чувствует себя удовлетворенным. Безосу приходит в голову идея использовать развивающийся интернет для создания сайта по продаже книг. Фильм наглядно показывает, насколько важно управление данными для электронной коммерции. Благодаря анализу пользовательских данных Amazon превратился из обычного сайта в цифровой конгломерат, известный по всему миру, в частности, из-за использования рекомендательных алгоритмов и прогнозирования спроса на основании обработанных клиентских данных.

Источник

Большие данные

2023

«Большие данные» – это образовательный проект при поддержке факультета компьютерных наук НИУ ВШЭ, в рамках которого выпускаются лекции, посвященные искусственному интеллекту, опасности цифровых «развлечений» и иным неблагоприятным аспектам для обычных пользователей при работе с большими данными.

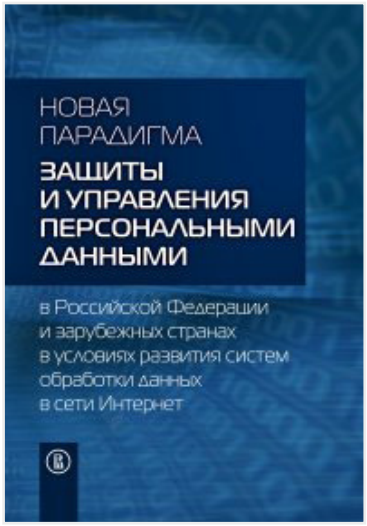
В настоящее время на федеральном канале НТВ вышло три эпизода.

- Первый, «Никто не избежит», рассказывает о том, как собираются данные и почему в результате их обработки возможно осуществлять какие-либо прогнозы для различных областей экономики и науки.
- Второй, «Цифровая пандемия», посвящен обработке больших данных искусственным интеллектом в медицине: эксперты рассказывают, как строится процесс обработки больших данных и как искусственный интеллект может на основе такой обработки по имеющимся симптомам «предсказать» диагноз пациента. Вопрос в том, насколько такие процессы безопасны для субъектов персональных данных и могут ли операторы обеспечить защиту их прав?
- Третий, «Цифровой рай», показывает пользователю, какие объемы своих данных он отдает «сети» при использовании цифровых сервисов и к каким последствиям для него это может привести.

Источник

ЧТО ПОЧИТАТЬ?

Рекомендации, статьи, книги



Источник

Книга «Новая парадигма защиты и управления персональными данными в Российской Федерации и зарубежных странах в условиях развития систем обработки данных в сети Интернет», Дупан Анна Сергеевна

Монография, изданная Анной Сергеевной Дупан при поддержке НИУ ВШЭ, представляет не только научный, но и особый практический интерес: в работе приводится сравнительный анализ подходов к управлению и защите персональных данных в странах разных юрисдикций в условиях применения новых методов обработки больших массивов данных и использования технологии облачных вычислений, которую можно заложить в основу для формирования «лучших практик» управления персональными данными.

Материалы форума DATA&AI 2024: на пороге экономики данных

Форум DATA&AI 2024 – важное событие года по теме больших данных. На панельных дискуссиях были рассмотрены ключевые темы в области управления данными, в том числе:

- Регулирование оборота данных: проблемы и перспективы.
- Безопасность данных: правила кибергигиены и актуальные подходы к защите.
- Искусственный интеллект как внешняя угроза и как внутренний страж.
- Data-driven бизнес: корпоративная стратегия и культура работы с данными.
- Национальный проект «Экономика данных» – новый этап взаимодействия государства и бизнеса в области данных и искусственного интеллекта.

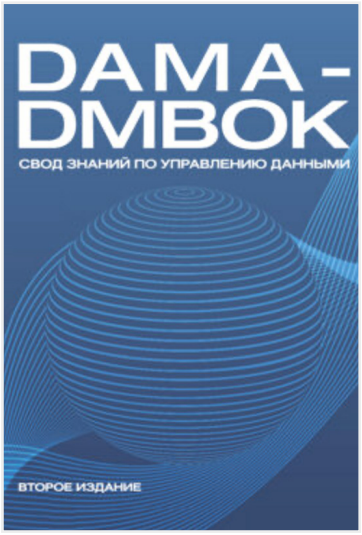
Форум полезен руководителям компаний и бизнес-подразделений, директорам по стратегии и инновациям, директорам по развитию и цифровой трансформации, ИТ-директорам и директорам по данным, бизнес-аналитикам, консультантам, исследователям и аналитикам данных, архитекторам и разработчикам информационных систем[1].

Если вы пропустили форум, с тезисами можно ознакомиться здесь, а с пресс-релизом тут. Официальный сайт форума доступен по ссылке (для ознакомления с материалами потребуется регистрация).

[1] DataAI 2024. URL: https://www.osp.ru/lp/data-ai2024 (дата обращения: 17.05.2024).

ЧТО ИЗУЧИТЬ?

Обучающие программы, онлайн-курсы, вебинары



Источник

DAMA-DMBOK2. Свод знаний по управлению данными. Кристофер Брэдли и коллектив авторов.

Если вы хотите глубже рассмотреть вопрос управления данными, изучить основы для внедрения практик управления данными в процессы своей компании и стать экспертом в этой области, рекомендуем познакомиться со сводом знаний по управлению данными DAMA DMBOOK и посмотреть интервью с Кристофером Брэдли – одним из его авторов.

Начинающим разбираться в этой теме можем порекомендовать для начала ознакомиться с кратким конспектом DAMA-DMBOK2, созданным в рамках Data Literacy Project, где в доступной форме раскрываются ключевые положения каждой из глав книги в формате удобного и простого для понимания конспекта.

Авторы

Андрей Никифоров

Эксперт в области персональных данных, команда DPO Блока B2C, Сбер

Олег Беляев

Руководитель направления, команда DPO Блока КИБ, Сбер

Яна Гришкова

Эксперт в области построения процессов кибербезопасности и приватности, команда DPO Блока Сеть продаж, Сбер

Екатерина Басниева

Эксперт в области приватности, дата-инженер, компания Группы Сбер

Алексей Булавин

Исполнительный директор управления развития технологий искусственного интеллекта и машинного обучения, SberData

PRIVACY-ПРОЕКТЫ КОМАНДЫ DPO СБЕРА

DATA PROTECTION OFFICE

DPO (Data Protection Office) – команда, которая отвечает за конфиденциальность и безопасность персональных данных в Сбере. Ежедневно DPO выполняет огромное количество задач, чтобы персональные данные клиентов, работников и партнеров Сбера были под надежной защитой. Наряду с основной деятельностью DPO запускает проекты, ориентированные на развитие цифровой грамотности населения. Мы хотим познакомить вас с ними.

Бесплатный онлайн-курс

ВСЕЛЕННАЯ ПЕРСОНАЛЬНЫХ ДАННЫХ

защита в интернете

Научитесь защищать личные данные вместе с нашим онлайн-курсом. Курс актуален для школьников, студентов, молодежи, взрослой аудитории, а также для особо незащищенных слоев населения – тех, кто находится на пенсии и наиболее часто становится жертвой мошенников.

Мы предлагаем нашим пользователям погрузиться во вселенную персональных данных и узнать:

- что такое персональные данные и как они могут попадать в интернет;
- как распознавать угрозы, которым подвержены личные данные, и противостоять им;
- как защищать свои права в эпоху киберугроз.

Пройти курс можно на платформе финансовой грамотности Сбера «СберСова». Чтобы начать изучение, необходимо войти на платформу с помощью Сбер ID – для этого достаточно указать номер телефона.

В конце каждого урока небольшой интерактивный тест. В финале – кибербитва. Это игра-симулятор на реальных кейсах. Успешно справившихся ждет почетный титул – Магистр вселенной персональных данных. Для тех, кто не смог отразить все угрозы, предусмотрены статусы Специалист и Новичок и возможность пройти кибербитву еще раз.

Рекомендуем данный курс для прохождения и использования в работе при обучении населения вопросам управления своими данными и защиты их от актуальных угроз.

PRIVACY-ПРОЕКТЫ КОМАНДЫ DPO СБЕРА

DATA PROTECTION OFFICE

DPO (Data Protection Office) – команда, которая отвечает за конфиденциальность и безопасность персональных данных в Сбере. Ежедневно DPO выполняет огромное количество задач, чтобы персональные данные клиентов, работников и партнеров Сбера были под надежной защитой. Наряду с основной деятельностью DPO запускает проекты, ориентированные на развитие цифровой грамотности населения. Мы хотим познакомить вас с ними.

ОНЛАЙН-РЕСУРС SBER BANK PRIVACY

Sber Bank Privacy – ключевой ресурс Сбера, где мы рассказываем о том, как в Банке обрабатываются и защищаются персональные данные.

Проект ориентирован на то, чтобы сделать обработку персональных данных прозрачной и донести до клиентов: защита персональных данных – ключевой приоритет Сбера, передавать данные в Банк – надежно и безопасно.

Для экспертов в области персональных данных на Sber Bank Privacy реализован отдельный раздел, в котором можно найти рекомендации по построению эффективной системы организации обработки и защиты персональных данных, чек-листы, которые помогут пройти проверки регулятора, профессиональный журнал команды DPO о приватности и безопасности персональных данных.

Заходите к нам,
мы будем вам рады

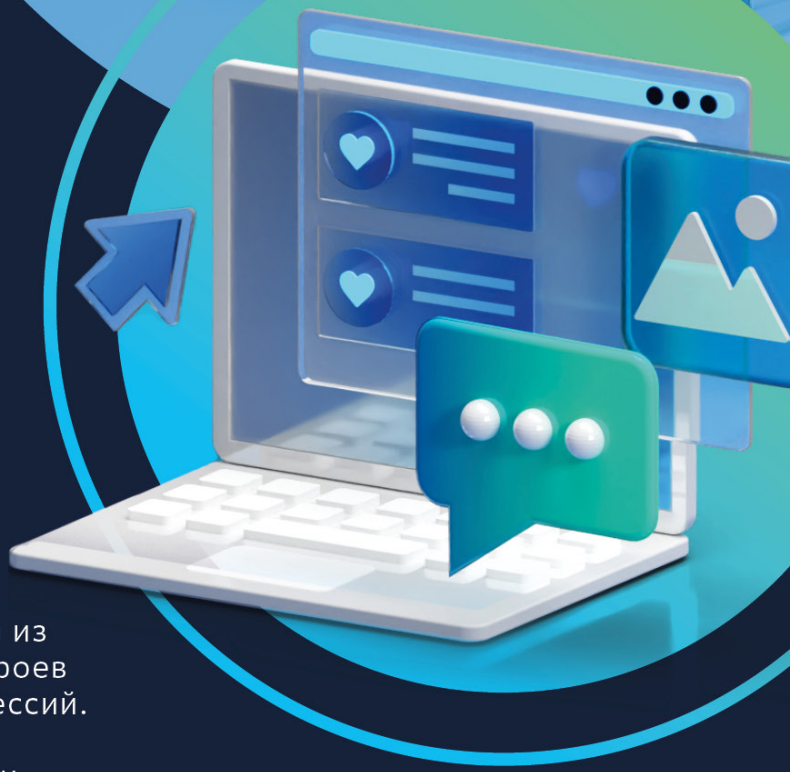
PRIVACY-ПРОЕКТЫ КОМАНДЫ DPO СБЕРА

DATA PROTECTION OFFICE

DPO (Data Protection Office) – команда, которая отвечает за конфиденциальность и безопасность персональных данных в Сбере. Ежедневно DPO выполняет огромное количество задач, чтобы персональные данные клиентов, работников и партнеров Сбера были под надежной защитой. Наряду с основной деятельностью DPO запускает проекты, ориентированные на развитие цифровой грамотности населения. Мы хотим познакомить вас с ними.

СТОРИТЕЙЛЫ ИЗ СЕРИИ «DPO НА СВЯЗИ»

истории из реальной жизни с рекомендациями по защите личных данных



«DPO на связи» – это сторителл-проект, в котором мы делимся реальными историями из жизни. Каждый рассказ строится вокруг героев – людей разных возрастов, статусов, профессий. Они попадают в ситуации, где их личные данные и они сами оказываются в опасности. Мы рассказываем, почему так произошло, что делать в таких обстоятельствах и как не попасть в неприятности снова. Возможно, в каких-то историях вы узнаете себя или своих близких.

Новые материалы выходят каждые две недели.

Читайте нас на платформе финансовой грамотности Сбера «СберСова», делитесь историями, рекомендуйте! Если материал вам понравился, авторизуйтесь по SberID и поставьте нам лайк, ответив на вопрос в конце истории: «Вам понравилась статья?». Ссылки на истории:

История № 1
Как избежать утечки персональных данных

История № 2
Как распознать уловки мошенников

Все истории

ЭФФЕКТИВНОЕ ОБРАЗОВАНИЕ

2023

CYBER MEDIA

ПРЕМИЯ «КИБЕРПРОСВЕТ»

2023

SBER PRIVACY JOURNAL

СБЕР КИБЕР БЕЗОПАСНОСТЬ

ВЫПУСК №9 | ИЮНЬ 2024