# Книга о качестве данных

Сергей Дегтев

Сергей Дегтев

# Книга о качестве данных

«Автор»
2026

**Дегтев С.**

Книга о качестве данных / С. Дегтев — «Автор», 2026

Что такое качество данных за 30 секундДанные чистят не ради перфекционизма, а чтобы бизнес не терял деньги.Семь измерений качества: СВУПК+Д (Своевременность, Валидность, Уникальность, Полнота, Консистентность, Доступность + Точность).Автоматизируйте правило контроль  триггер (ПКТ), а не чините вручную.Одна ошибка в таблице может стоить миллиона. Дубликаты, NULL, устаревшие цены, неверный формат - это не досадные мелочи, это реальные потери.В этой книге - полный набор инструментов для работы с качеством данных: от простых SQL-проверок до автоматизации с Great Expectations, наблюдаемости, Data Contracts и этики.

# Содержание

# Сергей Дегтев
# Книга о качестве данных

## ВВЕДЕНИЕ

### О том, как написана эта книга

Добро пожаловать.

Прежде чем мы перейдём к метрикам, пайплайнам и правилам, которые экономят бизнесу время и деньги, важно честно рассказать, как появился этот текст.

Эту книгу создал не один человек в одиночку. Она родилась в тандеме живого автора-куратора и большой языковой модели. Мы работали по принципу «архитектор + строитель»:

· Человек задавал вектор, формулировал промпты, оттачивал метафоры, придумывал запоминающиеся акронимы (ЧПУ, НДУНФ, СВУПК+Д и другие) и собирал разрозненные фрагменты в единую логическую цепочку.

· ИИ выступал в роли неутомимого соавтора: генерировал черновики глав, писал примеры кода, составлял таблицы, чек-листы и формулировал «якорные фразы».

Такой подход позволил упаковать многолетний отраслевой опыт в компактный, структурированный и сразу применимый формат.

### Но у него есть важное ограничение

ИИ не обладает интуицией, может ошибаться в технических деталях или использовать устаревшие синтаксисы. Поэтому автор (человек) не несёт ответственности за возможные неточности, «галлюцинации» модели или изменения в инструментах. Все примеры кода - это рабочие шаблоны, а не готовые production-решения. Перед внедрением обязательно адаптируйте и протестируйте их в вашей среде.

### Юридическое предупреждение

Всё, что описано в этой книге, - методики, код, шаблоны, чек-листы, рекомендации - предоставляется «как есть», без каких-либо гарантий. Вы используете их на свой страх и риск.

Автор не несет ответственности за любые прямые или косвенные убытки, потерю данных, финансовые потери, репутационные риски или юридические последствия, возникшие в результате применения материалов книги.

Перед внедрением описанных подходов в вашей компании обязательно проконсультируйтесь с квалифицированными юристами, специалистами по комплаенсу и безопасности, а также сверьте предлагаемые решения с актуальным законодательством (включая, но не ограничиваясь, 152-ФЗ, GDPR, отраслевыми стандартами).

Автор не гарантирует, что примеры кода будут работать в вашем окружении без доработок, а также что упомянутые инструменты и сервисы сохранят свои свойства на момент вашего прочтения.

# Почему качество данных - это отдельная дисциплина

### Метафора

Представьте, что 100 лет назад вы строите дом. Вам нужны: плотник, каменщик, кровельщик. Но вам не нужен отдельный специалист по качеству гвоздей.

Почему? Потому что:
· Гвоздей мало (десяток коробок на дом)
· Их можно пересчитать вручную
· Если гвоздь кривой - это видно сразу
· Ошибка стоит недорого (заменил гвоздь - и дальше работаешь)

А теперь представьте, что вы строите небоскрёб. Вам нужны миллионы крепёжных элементов. Вы не можете проверить каждый. Ошибка в одном типе болта может обрушить этаж. И вы вынуждены нанять отдельного человека, который знает стандарты, проверяет сертификаты, тестирует выборку и отслеживает, откуда пришли детали.

Качество данных - это тот же «специалист по болтам». Он появился не потому, что люди стали ленивее. А потому, что масштаб, скорость и цена ошибки выросли в миллионы раз.

### Определение

**Качество данных** - это степень их пригодности для решения конкретной бизнес-задачи. Данные не бывают «чистыми» или «грязными» в абсолютном смысле. Они бывают достаточно хорошими для прогноза оттока, но непригодными для налоговой отчётности.

## Как менялись данные и почему контроль перестал быть ручным

### Эпоха перфокарт (1960–1980)

Данных было мало. Их можно было проверить глазами. За качество отвечал программист, который их и вводил.

### Эпоха баз данных (1990–2010)

Данных стало больше. Появились CHECK, FOREIGN KEY. База начала сама «не пускать» очевидный мусор. За качество отвечали администраторы и аналитики.

### Эпоха Big Data и ИИ (2010–сейчас)

Миллиарды записей, сотни источников, решения в реальном времени. Здесь человеческий контроль уже не работает. Появилась роль Data Quality Engineer.

## Пять причин, почему DQ выделилось в отдельное направление

**1. Масштаб**

Раньше - 10 000 строк в Excel. Сейчас - 10 миллионов событий в час. Никакой человек не проверит это вручную. Нужны автоматические конвейеры.

**2. Скорость**

Раньше отчёт строился раз в месяц. Сейчас кредитный скоринг, фрод-детекция и персонализация работают в реальном времени. Ошибка стоит денег здесь и сейчас.

**3. Разнообразие источников**

Раньше данные вводились в одну систему. Сейчас они летят из CRM, ERP, мобильных приложений, IoT-датчиков и внешних API. Каждый источник - свой формат и свои ошибки.

**4. Регуляторика.**

Раньше «ошиблись в отчёте - пересадим». Сейчас 152-ФЗ, GDPR, отчётность в ЦБ и ФНС. Ошибка в данных - это штраф до 4% от оборота или блокировка счёта.

**5. ИИ и ML.**

Раньше данные шли в отчёты. Сейчас они кормят модели. Ошибка в обучающей выборке означает, что модель выучит неправильные паттерны и будет ошибаться автоматически и уверенно.

Эти пять причин не возникли в один день. Они накапливались десятилетиями. Вот как это происходило - краткая эволюция DQ.

## Короткая история профессии

DQ выросло из практики. В 1990-х появились хранилища (DWH) - и данные начали «портиться» при переносе. В 2000-х закон SOX в США заставил компании юридически отвечать за точность финансов. В 2010-х Big Data сделал ручной контроль невозможным. В 2015-м GDPR превратил качество из «хорошо бы» в юридическое требование.

А сегодня генеративный ИИ усилил **принцип GIGO:** мусор на входе теперь порождает не просто ошибку, а уверенную галлюцинацию.

**Что изменилось для вас лично?**

Раньше вы могли быть «просто аналитиком» или «просто инженером», и качество данных было побочной задачей. Сейчас, если вы работаете с данными, вы автоматически отвечаете за их пригодность. Вопрос только в том, будете ли вы делать это осознанно - с инструментами и процессами - или продолжите «тушить пожары» вручную.

**Почему эта книга написана сейчас?**

Не потому, что DQ стало «модным». А потому, что практика опередила литературу:

DQ вышло из IT-подвала.

Это больше не техническая задача «админа базы». Это бизнес-дисциплина, которая напрямую влияет на выручку, комплаенс и устойчивость ИИ-моделей.

**Инструменты готовы к продакшену.**

Great Expectations, dbt, Data Contracts - это не эксперименты, а рабочий стек. Но знания о них размазаны по официальной документации, форумам и разрозненным статьям.

Между академикой и практикой - пустота.

Когда я сам искал материал, выбор был таким: либо 600-страничные энциклопедии стандартов (вроде DAMA DMBOK), либо фрагментарные посты без связной системы. Не хватало одного компактного руководства, которое соединяет теорию, код и реальные процессы без воды, пустых обещаний и попытки объять необъятное.

Эта книга - попытка заполнить этот пробел. Практический минимум для тех, кому нужно начать настраивать проверки, считать ROI и выстраивать процессы завтра, а не через полгода изучения теории.

**Якорная фраза**

*Качество данных выделилось в отдельное направление не потому, что люди стали ленивее. А потому, что масштаб, скорость и цена ошибки выросли в миллионы раз. DQ - это не мода. Это ответ на объективную реальность.*

Вы спросите: «Хорошо, я понял, что качество данных – это отдельная дисциплина. Но зачем мне, обычному аналитику или инженеру, в это вникать? У меня и так отчёты горят».

Затем, чтобы не быть голословным, перейдём к самому убедительному аргументу – деньгам. Одна ошибка в таблице может стоить миллионов. И сейчас вы увидите это на цифрах.

# Почему качество данных - это деньги

Перед тем как нырять в технические детали, давайте договоримся о главном. Качество данных - это не про «нравится / не нравится». Это про деньги. Одна ошибка в таблице может стоить миллионов. Прочитайте эту главу - и вы никогда не скажете «подумаешь, опечатка».

**Метафора**

Представьте, что вы курьер в доставке еды. Навигатор показывает адрес, но координаты сбиты на 200 метров. Вы едете не туда. Клиент не получает пиццу. Компания теряет 500 рублей. За день так 10 раз. За месяц - 150 000 рублей. За год - почти 2 миллиона. И никто специально не врал. Просто данные чуть-чуть ошиблись.

**Качество данных** - это не про перфекционизм. Это про деньги, время и доверие.

**Важное уточнение:** чистота не равно полезность

Перед тем как мы перейдём к страшным историям о потерянных деньгах, давайте зафиксируем одну тонкость, о которой молчат 90% статей по качеству данных.

Можно добиться 100% полноты, уникальности и валидности. Можно убрать все NULL, все дубликаты, все неверные форматы. И при этом данные останутся бесполезными для принятия бизнес-решений. Почему? Потому что вы чистили не те поля, не с той тщательностью и не для той задачи.

**Пример**

Команда полгода наводила порядок в справочнике «Товары»: убрала дубликаты, проставила все категории, унифицировала единицы измерения. Гордость отдела качества данных. Приходит аналитик и говорит: «Мне для прогноза продаж нужна история цен за последние 3 года. А у вас хранится только текущая цена. Ваш чистый справочник бесполезен».

Технически чистые данные могут быть бизнесово бесполезны. Полезные данные могут быть технически грязными, но достаточными для решения.

**Вывод**

Чистота ради чистоты - это игрушки. Полезность для решения - это деньги. Перед тем как чистить данные, спросите себя: «Какое бизнес-решение будет принято на основе этих данных?» Если ответа нет - возможно, вы чистите не то.

**Экономическая модель COPQ** (Cost of Poor Data Quality)

**Формула стоимости низкого качества (COPQ)**

*Прямые потери* (ошибочные отгрузки, штрафы) + *косвенные* (человеко-часы на ручную чистку) + *репутационные* (отток клиентов, снижение доверия).

Если вы не можете посчитать хотя бы два из трёх слагаемых — вы не готовы защищать проект перед бизнесом.

**Три страшных истории из реальной жизни**

| *Ситуация* | *Цена ошибки* | *Почему произошло* |
|---|---|---|
| *Два одинаковых заказа → отгрузка двух холодильников* | *−60 000 ₽* | *Дубликат в таблице заказов* |
| *Дата рождения клиента 01.01.1900 → кредитный скоринг сломался* | *Потерян клиент на 2 млн ₽* | *Невалидная (неправильная) дата - заглушка вместо пустого значения* |
| *В отчёте продажи за октябрь = 0 → менеджеры уволились* | *Хаос в управлении* | *Процесс переноса данных не загрузил информацию 3 дня* |

Что общего у этих трёх историй? Везде входные данные были чуть-чуть не те - а последствия выросли до миллионов. Их объединяет один и тот же механизм - закон GIGO. Давайте разберёмся, почему так происходит и как с этим жить.

**Принцип GIGO (Garbage In, Garbage Out)**

Вы наверняка слышали эту фразу. Но давайте честно: в неё верят, когда речь о чужом коде, и забывают, когда речь о своих данных.

GIGO расшифровывается как
**G**arbage **I**n, **G**arbage **O**ut - «мусор на входе, мусор на выходе».

**Что это значит**
Если в вашу систему попали грязные данные - на выходе вы получите грязные отчёты, предсказания и решения. Система не может «додумать» чистоту. Она честно переработает мусор в мусор.

**Три уровня GIGO (как это выглядит в реальности)**

| ***Уровень*** | ***Что попало на вход*** | ***Что получили на выходе*** | ***Во сколько обошлось*** |
|---|---|---|---|
| *Смешно* | *В поле «возраст» ввели 999* | *Аналитик полдня искал «клиента-вампира»* | *4 часа рабочего времени* |
| *Больно* | *Дубликат заказа из-за двойного нажатия кнопки* | *Отгрузили два холодильника вместо одного* | *60 000 ₽* |
| *Смертельно* | *В обучающую выборку ML-модели попали неправильные метки* | *Модель предсказывает кредитный рейтинг со смещением* | *Потерянный клиент на 2 млн ₽ + репутационные риски* |

**Почему GIGO - это не просто поговорка, а закон**

# Конец ознакомительного фрагмента.