УДК 004.62+ 004.04

ОЦЕНКА КАЧЕСТВА БОЛЬШИХ ДАННЫХ

Часть 1. Основные понятия и метрики

Бурый А.С., д-р техн. наук, ФГБУ «Институт стандартизации»

Погодин И.М., аспирант ФГБУ «Институт стандартизации»

Концепция больших данных стала общеизвестной из-за широкого распространения информационно-коммуникационных технологий, Интернета вещей, технологии 5G, облачных сервисов и ряда других. Большие данные – это не только данные, но и полный концептуальный и технологический инструментарий, включающий необработанные и обработанные данные, хранилища для них, способы управления данными, модели и методы обработки и анализа данных. Целью исследования является разработка концептуального подхода к оценке качества больших данных исходя из сложившихся требований к качеству обычных данных, отмечаются существенные особенности рассмотрения признаков качества с учетом внутренних, контекстных и технологических особенностей больших данных. На основе анализа частных показателей качества данных представлены информационные особенности: источники, признаки получения больших данных, а также концептуальные подходы к интеграции разнообразных данных, включая ETL-процессы.

Ключевые слова: данные, качество данных, процессы анализа данных, интеграция данных, цифровая платформа, контроль качества данных, ETL-процесс, аспекты качества больших данных.

Цитирование: Бурый А.С., Погодин И.М. Оценка качества больших данных. Часть 1. Основные понятия и метрики // Информационно-экономические аспекты стандартизации и технического регулирования. 2024. № 3 (78). С. 49–58.

ВВЕДЕНИЕ

С появлением новых технологий и методов сбора данных возникает необходимость не только в их обработке, но и в обеспечении высокого уровня их точности, достоверности и целостности. В этом контексте тема автоматизации процессов обеспечения качества данных, значительная часть которых не структурирована, требует дополнительных усилий по обработке, структуризации, преобразованию потока сырых данных в информацию, полезную в практической деятельности [1].

Активное использование систем беспроводной связи, электронных и мобильных устройств, а также сетевых коммуникаций обеспечивает постоянный поток данных, генерируемых пользователями и техническими объектами в режиме реального времени об их местоположении, состоянии (физическом и техническом) – это лишь некоторые из источников, формирующих информационные потоки, получившие название больших данных. Сегодня концепция больших данных (BD – Big Data) стала общеизвестной из-за широкого распространения технологий (например, Интернета вещей, технологии 5G), а средства массовой информации почти ежедневно напоминают нам о ее важности для развития компаний и организаций. Концепция цифровой трансформации [2] заставляет нас по-новому посмотреть на интеграцию таких технологий, как облачные вычисления, большие данные, искусственный интеллект, Интернет вещей, которые уже сейчас обеспечивают конкурентные преимущества многим компаниям.

Оперативная ориентация в информационном пространстве больших данных, при использовании алгоритов интеллектуального анализа данных позволяет в сжатые сроки получать необходимую информацию для управления, анализа рынков, изучать тенденции спроса и предложений [3], расширяя области применения, включая государственные информационные системы [4] в рамках систем межведомственного электронного документооборота [5].

Традиционно работа с массивами данных сводилась к разработке баз данных (БД), основными требованиями к которым были: логическая структура, удобство и простота поиска информации, осуществление предобработки средствами системы управления БД [6]. Для распределенных систем поддержки и принятия решений модели хранения данных строятся под конкретную предметную область, а для междисциплинарного взаимодействия в модели хранения данных предусматривается унифицированный информационный интерфейс [7].

Умение работать с BD, на которых строятся краткосрочные и долгосрочные стратегии планирования и управления, аналитика производственной деятельности во многом определяет эффективность современных организаций, результативность научных исследований. Качество данных (КД) – характеристика, которая отражает степень их пригодности к использованию. В зависимости от сферы использования это понятие может относиться и к множеству значений количественных либо качественных переменных.

Целью исследования является разработка концептуального подхода к оценке качества больших данных исходя из сложившихся требований к качеству обычных данных и существенных особенностей рассмотрения признаков качества с учетом внутренних, контекстных и технологических особенностей больших данных.

СОСТАВЛЯЮЩИЕ КАЧЕСТВА ДАННЫХ

Эффективность применения и качество функционирования производственных объектов во многом определяются качеством данных, на основе которых принимаются управленческие решения. Инструменты доступа к данным должны сочетаться с корпоративными технологиями управления данными с целью совершенствования их качества [8].

Таблица 1

Проблемы с качеством данных и производные требования

ПРЕДМЕТНАЯ ОБЛАСТЬ	ПРОБЛЕМЫ С ДАННЫМИ	ПРОБЛЕМЫ С УПРАВЛЕНИЕМ КАЧЕСТВОМ ДАННЫХ	ТРЕБОВАНИЯ К КАЧЕСТВУ ПРОИЗВОДНЫХ ДАННЫХ
Номенклатурные данные	Дублирование; недостающие данные; нерелевантные данные	Проверка данных о товарах; управление нерелевантными данными; время реакции; анализ влияния на бизнес	Достоверность; безопасность; доступность
Данные об организации	Дублирование; не относящиеся к делу данные; отсутствуют справочные данные	Измерение качества данных; соглашения об уровне обслуживания (SLA); стандарты обработки данных; справочные основные данные; управление данными	Достоверность; актуальность; репутация; представление; дополнительная ценность
Данные о людях (сотрудниках, клиентах)	Неполные данные; неактуальные данные	Анализ первопричин; протоколы для решения проблем; понимание политик; права собственности на данные; технология; проблемы с отслеживанием	Точность; безопасность; дополнительная ценность; своевременность; актуальность
Данные об услугах/активах	Отсутствующие данные; неполные данные; неверные значения; устаревшие данные; дублирование	Требования; метрики; измерения; стандарты; анализ первопричин; анализ влияния на бизнес	Доступность; дополнительная ценность; полнота; своевременность; интерпретируемость
Управление цепочкой поставок	Ошибки проектирования; дублирование; проблемы с основными данными	Метрики; измерения; отчеты/протоколы; соглашения об уровне обслуживания; управление данными; наблюдаемость данных	Доступность; безопасность; своевременность; точность; полнота

Основным стандартом в области качества данных является ГОСТ Р ИСО 8000-2–2019: «Качество данных». Ч. 2[1], где под качеством данных понимается степень, с которой характеристики данных удовлетворяют заявленным требованиям при использовании в заданных условиях.

Стандарт является частью серии международных стандартов ИСО 8000, посвященных качеству данных. Он представляет собой обзор комплекса стандартов ИСО 8000, рассматривающих вопросы качества основных данных.

Основные данные – это данные, которые используются для описания сущностей в рамках всей организации. В зависимости от предметной области (ПрО) используемых данных к КД применяются различные требования, представленные в табл. 1, составленной с учетом [9].

Проблемы управления качеством данных в различных областях данных показывают, что одни компании оказываются более продвинутыми в вопросах управления качеством данных, чем другие [9]. Тем не менее, в каждой ПрО есть возможности для совершенствования. Так, для данных о физических лицах основные проблемы связаны с обеспечением достоверности данных. Несмотря на то, что общие возможности для измерения КД не идеальны, в некоторых областях полностью отсутствуют средства измерения КД и отчетности, включая соглашения об уровне обслуживания для обеспечения качества данных, управление данными и политика в области КД, установка стандартов обработки данных и бизнес-правила.

Основные источники некачественных данных для каждой предметной области могут быть свои (см. табл. 2), включая случайные и преднамеренные факторы.

Некачественными данными признаются данные, характеристики которых не соответствуют и (или) не достигают установленных для них требований, что делает невозможным для пользователей данных полагаться на них, тогда как данные с высоким качеством – это надежная база для любого процесса и операции, способствующая «правиль-

[1] ГОСТ Р ИСО 8000-100–2019. Качество данных. Часть 2. Словарь. – М.: Стандартинформ, 2019. – 11 с., IV (Введ. 2020-05-01).

Таблица 2

Источники некачественных данных (составлено с учетом [10])

№		ИСТОЧНИК	КРАТКОЕ СОДЕРЖАНИЕ
1	«Человеческий фактор»	1.1 Действия персонала	Неумышленные действия персонала, связанные с квалификацией, опытом, состоянием и т.д.
		1.2 Эксплуатация программных средств (ПрС)	Использование непроверенных данных или нелицензионного программного обеспечения (ПО)
		1.3 При разработке ПО	Ошибки в обеспечении безопасности при разработке, эксплуатации, сопровождении ПО
2	Социальный инжиниринг		Умышленные действия сторонних лиц, преследующих мошеннические цели
3	Несанкционированный доступ	3.1 Физический	Доступ неавторизованных лиц (ДНЛ) в контролируемую зону расположения аппаратных и ПрС
		3.2 Логический	ДНЛ к информационным активам (компрометация пароля, ошибки администрирования и др.)
4	Организационные факторы	4.1 Нарушения сторонними (третьими) лицами обязательств	Связано с поставкой некачественного ПО и (или) оказания услуг
		4.2 Для обеспечения безопасности на стадиях жизненного цикла	Нарушения в организации безопасности как при разработке, так и в ходе эксплуатации информационной системы

ной» интерпретации информации и данных и их целевому использованию [10].

Для управления процессами взаимодействия (на межмашинном уровне – M2M, человеко-машинном уровне – P2M и др.) в ходе решения задач интеграции информационно-коммуникационных технологий (ИКТ) основным требованием выступает обеспечение согласованной координации данных, информации и знаний, а также надежной адаптации информационных массивов, ориентированной на максимизацию информационной эффективности [4]. Для этого предлагается использовать методику «двунаправленного вычислительного моста» между указанными сферами, основанную на пирамиде знаний Р. Акоффа[2].

Пирамида трансформации данных от этапа их получения до целевого применения знаний лицами, принимающими решение (ЛПР), представлена на рис. 1. Задачи обеспечения качества сначала данных, потом получаемой информации и т. д. последовательно решаются в соответствии с этапами трансформации данных. Аналитики или инженеры по данным понимают, как получить данные, как использовать их контекст, текущий формат, как извлечь из данных полезную информацию; специалисты по исследованию данных (data scientist) могут обнаружить скрытые закономерности, подозрительные аномалии или логические связи в данных [8].

Рис. 1. Пирамида трансформации данных

За оценку качества данных отвечают инженеры Data Quality (инженеры по КД – ИКД), которые управляют информационными массивами, проверяют их поведение в текущих и новых условиях, контролируют релевантность, достаточность и актуальность. Как правило, обязанности ИКД не ограничиваются только рутинными проверками записей в таблицах систем управления БД, а требуют глубокого понимания бизнес-потребностей, чтобы трансформировать имеющиеся данные в пригодную к практическому использованию информацию.

В соответствии с ГОСТом и другими стандартами процессы управления КД делятся на три группы: выполнение операций над данными, непрерывный контроль качества данных, повышение КД.

Инструменты управления качеством данных. У бизнеса есть широкий арсенал инструментальных систем для обработки и выполнения операций с данными. К ним относятся системы управления мастер-данными, продукты Data Quality, программы для работы с аналитикой. Процесс управления КД для любой организации является уникальным исходя из ее целевых задач. Однако основные принципы управления включают (рис. 2): организационный аспект (модели данных, разработка процедур контроля, описание потоков и др.); структуризацию данных (профилирование); информационную безопасность (обеспечение конфиденциальности и устойчивости данных, описания инцидентов и способов реагирования на них); аналитику данных, включая контроль, мониторинг, анализ качества; контекстуальность данных в зависимости от ПрО.

Рис. 2. Составляющие управления качеством данных

Показатели качества данных условно – критерии можно разбить на 3 группы. Первая группа – это требования к содержанию данных (внутренние свойства), вторая – согласованность информации, а третья – удобство (технологичность) работы.

Наиболее критичны проблемы первой группы. Их игнорирование может привести к полной невозможности анализировать. Отсутствие согласованности снижает доверие к принимаемым решениям, а неудобство использования по-

[2] Ackoff R.L. From data to wisdom // Journal of Applied Systems Analysis. 1989. Vol. 16. P. 3–9.

вышает затраты. В табл. 3 представлены основные показатели качества в зависимости от групп КД.

Анализ интерпретации показателей качества показал, что определение характеристик в различных источниках не

Таблица 3

Составляющие качества данных

№ П/П	ПРИНЦИПЫ ГРУППИРОВАНИЯ	СОСТАВЛЯЮЩИЕ ПОКАЗАТЕЛЯ КАЧЕСТВА
1	Требования к содержанию данных	1.1 Достоверность (accuracy) – это соответствие данных реальности и корректность их интерпретации. Неточные данные нельзя использовать для принятия решений
		1.2 Полнота (completeness) – достаточность объема, глубины и широты наборов данных. Неполнота приводит к невозможности анализа
		1.3 Релевантность (relevance) – показатель того, насколько данные соответствуют целям и решаемым задачам
		1.4 Объективность (objectivity) – уверенность, что данные не содержат предвзятых мнений или субъективных оценок
		1.5 Валидность (validity) – соответствие многочисленным атрибутам, связанным с элементом данных: тип, точность, формат, диапазоны допустимых значений и т. д.
		1.6 Точность (precision) – детальность измерения и фиксации данных
		1.7 Своевременность (timeliness) – время после сбора данных, по истечении которого они становятся доступными для анализа. Корректные, но устаревшие данные бесполезны для принятия оперативных решений
2	Согласованность информации	2.1 Уникальность (uniqueness) подразумевает единственность объекта в наборе данных. Дублирование данных может приводить к несогласованности и противоречиям
		2.2 Целостность (integrity) – наличие корректных ссылок между данными и их соответствие установленным правилам и ограничениям
		2.3 Согласованность (consistency) – соответствие данных друг другу и их логическая непротиворечивость. Несогласованность данных указывает на ошибки или неточности в их сборе или обработке
		2.4 Когерентность (coherence) – согласованность с другими источниками данных и логикой процесса, который они описывают
		2.5 Надежность (reliability) – возможность повторного получения одинаковых результатов
3	Удобство использования	3.1 Доступность (accessibility) показывает, насколько легко пользователю узнать, какие данные имеются в его распоряжении, а также получить доступ к ним. Причем речь может идти в том числе и о метаданных, описывающих анализируемую информацию
		3.2 Удобство использования (usability) характеризует, насколько легко и просто использовать данные для изучения определенной проблемы. Это особенно характерно для неструктурированных данных: изображения, аудио, видео
		3.3 Универсальность (universality) определяет, насколько данные могут использоваться для разных целей и задач.
		3.4 Контролируемость (traceability) или прослеживаемость – возможность осуществления контроля качества и происхождения данных с учетом источников, истории создания, изменения, преобразования, удаления, хранения и передачи
		3.5 Переносимость (portability) – возможность переноса данных между разными платформами или службами без потери их целостности. Сложности интеграции, импорта или экспорта данных существенно снижают их ценность

совпадает, что свидетельствует о еще неустоявшихся подходах к анализу признаков КД [11].

БОЛЬШИЕ ДАННЫЕ И ВОПРОСЫ КАЧЕСТВА

Наиболее известным определением BD является определение 3V, в котором говорится, что BD состоит из больших объемов данных (Volume), которые разнообразны (Variety) и поступают из разных источников и создаются с высокой скоростью (Velocity) [2]. Следует понимать, что BD – это не только данные, но и разрабатываемый аппаратно-программный комплекс, обеспечивающий сбор необработанных и обработанных данных, их хранение, способы управления данными, их обработку и аналитику. Задача, которая становится еще более сложной, – это управление качеством данных в средах больших объемов данных. Термин "большие данные" относится к структурированным или неструктурированным наборам данных [12], которые невозможно сохранить и обработать с помощью обычных программных средств (например, реляционных баз данных), независимо от вычислительной мощности или наличия физического хранилища. Как правило, объем означает размер данных, разнообразие означает неоднородность сбора данных, их представления и семантической интерпретации, скорость связана со скоростью предоставления (обновления) данных и временем, в течение которого данные являются актуальными для оперирования с ними. В прежние времена с большим объемом, например, измерительных данных «боролись», используя математические методы сжатия данных, что было эффективно для медленно меняющихся процессов [13]. Сегодня, когда объем данных растет по экспоненте: например, самолеты ежегодно генерируют 2,5 млрд ТБ данных с датчиков, установленных в двигателях[3], это может привести к невосполнимой потере данных.

Для извлечения ценности и повышения эффективности больших данных все чаще признается важность четвертого аспекта больших данных, то есть их достоверности. Достоверность (Veracity) напрямую связана с несоответствиями и проблемами качества данных: при огромном объеме генерируемых данных, высокой скорости их поступления и большом разнообразии разнородных данных КД далеко до совершенства. В табл. 4 характеристики BD расширены до 5V, хотя этот процесс продолжает развиваться уже до 7V и даже 10V.

Получение данных осуществляется с какой-либо целью. Данные нужны не ради данных. Образно их можно сравнить с куколкой, превращаемой в бабочку; так и в определенный момент данные преобразуются в знания (см. рис.1), пройдя ряд этапов преобразований, переработки, представленных на рис. 3. Методика извлечения знаний из баз данных, получившая название Knowledge Discovery Database (KDD) представляет собой набор атомарных операций, комбинируя которые можно получить нужное решение. Обычно этапы

[3] Формула Big Data: семь «V» + неординарная задача // Сайт «Форсайт». [Электронный ресурс]. URL: https://www.fsight.ru/blog/formula-big-data-sem-v-neordinarnaja-zadacha-2/ (дата обращения: 22.04.2024).

Таблица 4

Характеристики больших данных от 5 V, показатели описания V

№	BD, ПО V	ОПИСАНИЕ	АТРИБУТЫ И МЕТРИКИ
1	Объем (Volume)	Масштаб объема и размер данных в БД	От 1 ТБ (10^{12} Б) до 1ЭБ
2	Скорость (Velocity)	Частота генерации данных: скорость, с которой эти данные генерируются, обновляются и передаются в потоковом режиме	Пакетное, близкое по времени, реальное время распространения; постоянные потоки данных
3	Разнообразие (Variety)	Множество различных форм данных	Двоичные, текстовые, мультимедийные, структурированные, неструктурированные и полуструктурированные
4	Достоверность (Veracity)	Данные должны быть проверенными и достоверными, поскольку от исходных данных будет завесить результат анализа и качество принимаемых решений	Несогласованность, неполнота, двусмысленность, задержка, надежность и прослеживаемость (происхождение)
5	Ценность (Value)	Цель – извлечь максимум пользы из результатов анализа больших данных	Стратегия в области BD, целевые показатели и подходящий процесс аналитики

получения знаний включают выборку данных, их очистку, трансформацию, моделирование и интерпретацию полученных результатов [14].

Обеспечение высокого качества BD считается одним из наиболее сложных и критических этапов цепочки создания стоимости BD. Большинство существующих подходов применяют традиционные меры оценки КД (см. табл. 3) к BD. В [15] предлагается осуществлять оценку качества BD на основе комплексной системы оценки качества больших данных по 12 показателям: полнота, своевременность, изменчивость, уникальность, соответствие, согласованность, простота манипулирования, релевантность, читаемость, безопасность, доступность и целостность. Данные показатели группируются в 5 аспектов качества, а именно: актуальность, надежность, достоверность, доступность и удобство использования.

Каждый из этапов содержит ряд операций, определяемых ПрО и спецификой данных. Так для обеспечения качества BD наиболее актуальными являются этапы очистки и анализа данных, на которых производится профилирование данных, связанное с изучением данных от определенных источников данных, сбором/получением статистических данных, метаданных, установлением взаимосвязей, элементов управления данными и др. Профилирование может быть очень полезным при оценке полезности известных и неизвестных данных; это может помочь в развертывании будущих вариантов использования больших данных.

ЦИФРОВЫЕ ПЛАТФОРМЫ КАК «ГЕНЕРАТОРЫ» БОЛЬШИХ ДАННЫХ

Цифровая платформа – это основанная на совокупности технологий, продуктов и услуг информационная система, которая обеспечивает взаимодействие в единой (отраслевой, организационной) интернет-среде по заданным алгоритмам заинтересованных пользователей.

По сути цифровые платформы (ЦП) играют роль «оркестраторов», координируя информационное взаимодействие внутреннего бизнес-процесса предприятия с внутренними и внешними веб-сервисами в различных форматах – B2B, B2C, B2G, C2C и т.д.

Существуют следующие типы цифровых платформ[4], представленные на рис. 4:

- инструментальные ЦП, как совокупность аппаратных и ПрС для технологий обработки информации и данных, включая функционал для отладки прикладных программных или аппаратно-программных инструментов;

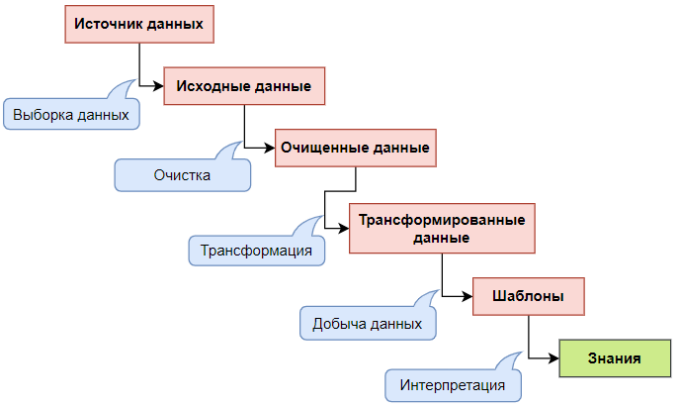

Рис. 3. Этапы процесса извлечения знаний

- инфраструктурные ЦП, связанные с экосистемами участников рынков информатизации, на которые выводятся ИТ-сервисы, использующие сквозные цифровые технологии работы с данными и доступ к источникам информации и применяемые в пределах экосистем[5];
- прикладные ЦП – бизнес-модели, обеспечивающие реализацию определенных алгоритмов обмена информацией и координации участников единого производственного цикла предоставляющие в различных ПрО [16].

Анализ научно-теоретических работ, посвященных проблеме исследования, позволил выделить основные виды цифровых платформ, представляющих интерес с точки рения развития и оптимизации деятельности социально-экономических систем различных уровней.

Полученные массивы больших данных, особенно если были задействованы различные ЦП, для дальнейшего использования и анализа необходимо интегрировать (объединить, представить в удобном формате, придав общую структуру). За объединение данных из ряда источников, очистку и упорядочение отвечают ETL-процессы. Аббревиатура ETL (от англ. Extract, Transform, Load) означает «извлечение, преобразование, загрузка», что и происходит с данными при их упорядочении. Обеспечение качества данных требует отслеживания дефектов качества в процессе ETL. Необходимость интеграции данных актуально в системах поддержки принятия решений и обусловлена такими причинами, как: разнородность форматов данных; устаревшие базы данных; изменение структуры источника данных с течением времени. Это делает КД неопределенным [17].

При интеграции сырых слабоструктурированных данных задачей ETL-процесса является извлечение сущностей

[4] Цифровые платформы // Сайт Центра развития компетенций в бизнес-информатике, логистике и управлении проектами (ВШЭ). [Электронный ресурс]. URL: https://hsbi.hse.ru/articles/tsifrovye-platformy/ (дата обращения: 22.03.2024).]

[5] Цифровая экосистема – это совокупность информационных систем (ЦП) различного функционала, чаще всего с общим интерфейсом. Экосистема обеспечивает клиентоцентричную бизнес-модель и объединяет участников, например, по разным категориям пользователей (бизнес, физические лица) с единой точкой входа (Яндекс, Сбер, VK и др.).

из исходных коллекций, их разрешение, трансформация и загрузка в интегрированное хранилище. Под сущностью здесь понимается некоторое цифровое представление объекта реального мира (например, информация о событии) [18]. При извлечении сущностей возникает проблема

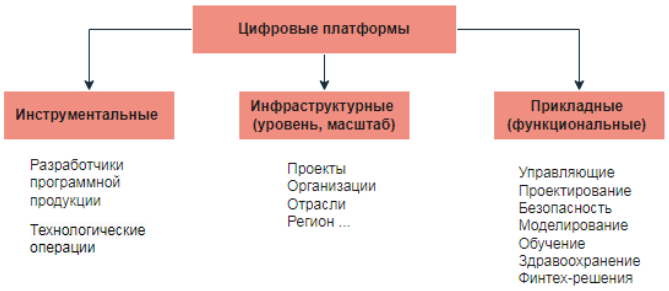

Рис. 4. Типы цифровых платформ

их разрешения: из разных ресурсов можно извлечь разную информацию об одном и том же объекте реального мира. Интегрированные потоки вида ETL выступают важным инструментом образования интегрированных структурированных качественных данных для дальнейшего анализа и обработки.

ЗАКЛЮЧЕНИЕ

Рассмотренные подходы к формированию качественных данных направлены на извлечение максимальной информации и знаний из данных для поддержки и принятия решений, координации взаимодействия на организационном, функциональном и технологическом уровнях управления предприятием и составления обоснованно успешных прогнозов.

Поскольку большие данные создают ценность не только с финансовой точки зрения, но и с точки зрения операционных (технологических) и стратегических преимуществ, изучение ценности BD и управление их качеством имеет решающее значение для успеха организаций и предприятий.

Во второй части данной работы планируется рассмотреть роль моделей представления данных и метрик для получения оценок качества данных, выявления закономерностей в данных, которая может послужить основой, как извлечения знаний, так и для организации хранения больших данных в интегрированных хранилищах данных.

Список использованных источников и литературы

1. Брюс П., Брюс Э. Практическая статистика для специалистов Data Science: Пер. с англ. – СПб.: БХВ-Петербург, 2019. – 304 с.
2. Сибел Т. Цифровая трансформация. Как выжить и преуспеть в новую эпоху. – М.: Манн, Иванов и Фербер, 2021. – 256 с.
3. Ючинсон К.С. Большие данные и законодательство о конкуренции // Право. Журнал Высшей школы экономики. 2017. № 1. С. 216–245.
4. Бурый А.С. Структуризация систем мониторинга информационных ресурсов // Правовая информатика. 2023. № 1. С. 52–61.
5. Бурый А.С., Слепынцева Л.И. Цифровизация контента документов по стандартизации. Часть 1. Состояние и современные тенденции // Информационно-экономические аспекты стандартизации и технического регулирования. 2021. № 1(59). С. 105–113.
6. Бурый А.С., Морин Е.В. Модельно-алгоритмические структуры оценки качества программных изделий. – М.: «Горячая линия-Телеком», 2019. – 160 с.
7. Единая модель хранения данных различных предметных областей для систем поддержки принятия решений / Ф.Г. Майтаков, А.А. Меркулов, Е.В. Петренко, А.Я. Яфасов // Морские интеллектуальные технологии. 2018. № 4–3 (42). С. 127–133.
8. Волков Д.В., Незнанов А.А. Качество данных: от стратегии к практике // Открытые системы. СУБД. 2020. № 1. С. 14–18.
9. Silvola R., et al. Data quality assessment and improvement // International Journal of Business Information Systems. 2016. Т. 22, №. 1. С. 62–81.
10. Савицкая М., Рощупкин И. Как оценить качество данных в информационных системах по Положению № 716-П и зачем это нужно // Внутренний контроль в кредитной организации. 2023. №1 (57). С. 66–81.
11. Бирюков А.Н. Качество данных как услуга // Прикладная информатика. 2020. Т. 15, № 4 (88). С. 120–132.
12. Firmani D., Mecella M., Scannapieco M., Batini C. On the meaningfulness of “big data quality” // Data Science and Engineering. 2016. № 1. С. 6–20.
13. Бурый А.С., Лобан А.В., Ловцов Д.А. Модели сжатия массивов измерительной информации в автоматизированной системе управления // Автоматика и телемеханика, 1998. № 5. С. 3–26.
14. Шелухин О.И., Ерохин С.Д., Ванюшина А.В. Классификация IP-трафика методами машинного обучения / Под ред. профессора О.И. Шелухина. – М.: Горячая линия – Телеком, 2020. – 284 с.

15. Elouataoui W., El Alaoui I., El Mendili S., Gahi Y. An advanced big data quality framework based on weighted metrics // Big Data and Cognitive Computing. 2022. No. 6 (4), 153.
16. Оборин М.С. Цифровые платформы как механизм рыночного взаимодействия продавцов и покупателей товаров и услуг // Sochi Journal of Economy. 2020. Т. 14, № 3. С. 292–301.
17. Souibgui M., Atigui F., Zammali S., Cherfi S., Yahia S.B. Data quality in ETL process: A preliminary study // Procedia Computer Science. 2019. Т. 159. С. 676–687.
18. Вовченко А.Е., Калиниченко Л.А., Ковалев Д.Ю. Методы разрешения сущностей и слияния данных в ETL-процессе и их реализация в среде Hadoop // Информатика и ее применения. 2014. Т. 8, № 4. С. 94–109.

ASSESSMENT THE QUALITY OF BIG DATA

Part 1. Basic concepts and metrics

Buryi A.S., Doctor of Sciences in Technology, Russian Standardization Institute

Pogodin I.M., graduate student of the Russian Standardization Institute

The concept of Big Data has become well-known due to the widespread use of information and communication technologies, the Internet of Things, 5G technology, cloud services and a number of others. Big Data is not only data, but a complete conceptual and technological toolkit, including raw and processed data, repositories for them, data management methods, models and methods of data processing and analysis. The aim of the study is to develop a conceptual approach to assessing the quality of Big Data, based on the prevailing requirements for the quality of ordinary data, noting the essential features of considering quality features taking into account the internal, contextual and technological features of Big Data. Based on the analysis of particular data quality indicators, information features are presented in the usual information system: sources, signs of obtaining Big Data, as well as conceptual approaches to integrating a variety of data, including ETL-processes.

Keywords: data, data quality, data analysis processes, data integration, digital platform, data quality control, ETL-process, aspects of Big Data quality.

For citation: Buryi A.S., Pogodin I.M. Assessment the quality of big data. Part 1. Basic concepts and metrics. Information and Economic Aspects of Standardization and Technical Regulation. 2024; 3 (78): 49–58. (In Russ.).

References

1. Bruce P., Bruce A. Practical statistics for data scientists, OREILLY Sebastopol, California–USA. 2017.
2. Siebel T.M. Digital transformation: survive and thrive in an era of mass extinction. Moscow. Mann, Ivanov i Ferber Publ., 2021. 256 p.
3. Yuchinson K.S. Bol'shie dannye i zakonodatel'stvo o konkurencii. Pravo. Zhurnal Vysshej shkoly ekonomiki, 2017, no. 1, pp. 216–245.
4. Buryi A.S. Strukturizaciya sistem monitoringa informacionnyh resursov. Pravovaya informatika, 2023, no. 1, pp. 52–61.
5. Buryi A.S., Slepynceva L.I. Cifrovizaciya kontenta dokumentov po standartizacii. Chast' 1. Sostoyanie i sovremennye tendencii. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya, 2021, no. 1 (59), pp. 105–113.
6. Buryi A.S., Morin E.V. Model'no-algoritmicheskie struktury ocenki kachestva programmnyh izdelij. Moscow: «Goryachaya liniya-Telekom» Publ., 2019, 160 p.
7. Majtakov F.G., Merkulov A.A., Petrenko E.V., YAfasov A.Y. Edinaya model' hraneniya dannyh razlichnyh predmetnyh oblastej dlya sistem podderzhki prinyatiya reshenij. Morskie intellektual'nye tekhnologii, 2018, no. 4–3 (42), pp. 127–133.
8. Volkov D.V., Neznanov A.A. Kachestvo dannyh: ot strategii k praktike. Otkrytye sistemy. SUBD, 2020, no. 1, pp. 14–18.

9. Silvola R., et al. Data quality assessment and improvement. International Journal of Business Information Systems, 2016, 22 (1), pp. 62–81.
10. Savickaya M., Roshchupkin I. Kak ocenit' kachestvo dannyh v informacionnyh sistemah po Polozheniyu № 716-P i zachem eto nuzhno. Vnutrennij kontrol' v kreditnoj organizacii, 2023, no. 1 (57), pp. 66–81.
11. Biryukov A.N. Kachestvo dannyh kak usluga. Prikladnaya informatika, 2020, vol. 15, no. 4 (88), pp. 120–132.
12. Firmani D., Mecella M., Scannapieco M., Batini C. On the meaningfulness of "big data quality". Data Science and Engineering, 2016, no. 1, pp. 6–20.
13. Buryi A.S., Loban A.V., Lovtsov D.A. Compression models for arrays of measurement data in an automatic control systems // Automation and Remote Control,1998, vol. 59, no. 5. Part 1, pp. 613–631.
14. Sheluhin O.I., Erohin S.D., Vanyushina A.V. Klassifikaciya IP-trafika metodami mashinnogo obucheniya. Pod red. professora O.I. Sheluhina. Moscow: «Goryachaya liniya – Telekom» Publ., 2020, 284 p.
15. Elouataoui W., El Alaoui I., El Mendili S., Gahi Y. An advanced big data quality framework based on weighted metrics. Big Data and Cognitive Computing, 2022, no. 6 (4), 153.
16. Oborin M.S. Cifrovye platformy kak mekhanizm rynochnogo vzaimodejstviya prodavcov i pokupatelej tovarov i uslug. Sochi Journal of Economy, 2020, vol. 14, no. 3, pp. 292–301.
17. Souibgui M., Atigui F., Zammali S., Cherfi S., Yahia S.B. Data quality in ETL-process: A preliminary study. Procedia Computer Science, 2019, vol. 159, pp. 676–687.
18. Vovchenko A.E., Kalinichenko L.A., Kovalev D.Yu. Metody razresheniya sushchnostej i sliya-niya dannyh v ETL-processe i ih realizaciya v srede Hadoop. Informatika i ee primeneniya, 2014, vol. 8, no. 4, pp. 94–109.Sidorov K.V., Rebrun I.A., et al. Diagnostika psihofiziologicheskogo i emocional'nogo sostoyaniya cheloveka-operatora. Inzhenernyj vestnik Dona, 2012, no. 4–2 (23), P. 27.