УДК 004.62+004.04

ОЦЕНКА КАЧЕСТВА БОЛЬШИХ ДАННЫХ. Часть 2. Модели данных

Бурый А.С., д-р техн. наук, Российский институт стандартизации

Погодин И.М., аспирант, Российский институт стандартизации

Концепция больших данных стала общеизвестной из-за широкого распространения информационно-коммуникационных технологий, сетевых технологий, облачных сервисов и ряда других. Продолжая исследования, начатые в части 1, уточнены виды моделей данных и их роль в анализе больших данных (BD).

Целью исследования второй части работы является анализ и применимость моделей данных в практических приложениях для моделирования процессов поддержки принятия решений и обоснованного выбора структуры данных для задач их комплексирования и интеллектуального анализа.Для задачи интеллектуального анализа BD на основе многоагентного подхода выявлены основные элементы, система отношений при формировании этапов переработки данных с целью обнаружения новых знаний.

Ключевые слова: большие данные, модель данных, показатели качества данных, интеллектуальный анализ данных, интеграция данных.

Цитирование: Бурый А.С., Погодин И.М. Оценка качества больших данных. Часть 2. Модели данных // Информационно-экономические аспекты стандартизации и технического регулирования. 2024. № 4 (79). С. 24–32.

ПРИНЯТЫЕ СОКРАЩЕНИЯ

БД – база данных
ИАД – интеллектуальный анализ данных
МД – модель данных
ПрО – предметная область
BD – большие данные (Big Data)

ВВЕДЕНИЕ

В первой части данной работы представлены основные источники генерации больших данных (BD), показаны этапы трансформации данных в информацию, а затем возможное извлечение знаний для поддержки и принятия решений, координации взаимодействия на организационном, функциональном и технологическом уровнях управления предприятием и составления обоснованно успешных прогнозов в зависимости от предметной области (ПрО) исследования [1].

Для формирования представления ПрО (объекты, отношения между ними) и выполнения операций в терминах таких представлений служат разнообразные реализованные в программных средах инструменты, например, компилирование, тестирование, моделирование. В рамках концепции моделирования основным функционалом выступают модели данных (МД), которые являются определяющими при разработке баз данных (БД). Не случайно в свое время создатель реляционной модели данных Э. Кодд стал лауреатом престижной Тьюринговской премии по информатике[1]. В современном понимании МД – это не результат, а основной механизм моделирования, т.е. совокупность правил структурирования данных, допустимых операций над ними и видов ограничений целостности [2].

К основным источникам BD в современной городской среде [3] можно отнести цифровую среду производственного сектора «умной» экономики [4] для аналитической поддержки применения процедур искусственного интеллекта, робототехнических устройств, приборов киберфизических систем и ряда других, а также социальную сферу. В рамках государственных информационных систем [5], за счет общедоступных панелей мониторинга, многочисленных информационных сервисов, пользователи могут выбирать интересующие их услуги, ориентироваться и участвовать в жизни города, непосредственно влиять, например, с позиции «активного гражданина», на улучшение городской среды [6].

Основная целевая задача любых данных заключается в возможности извлекать из них информацию, на основании, например, методов статистической обработки данных и моделирования [7], а также обнаруживать неявные, нео-

[1] Codd E.F. Relational database: A practical foundation for productivity // In ACM Turing award lectures. 2007, pp. 109–117.

чевидные закономерности в данных на основе методов интеллектуального анализа данных (ИАД), которые учитывать при принятии решений [8]. В последние годы, с интеллектуальным развитием промышленных предприятий, анализ BD становится основной движущей силой для предприятий, обеспечивая промышленную ценность, переводя промышленное производство в интеллектуальное русло. Производственные исследования, поддерживаемые данными, перешли к моделям, основанным на данных [9], а точнее – на BD. Уровень значимости больших данных для цифровой эпохи сегодня сравнивают с ролью нефти для общества, значимость которой становится все более весомой в результате ее полной переработки[2].

Целью исследования второй части работы является анализ и применимость моделей данных в практических приложениях для моделирования процессов поддержки принятия решений и обоснованного выбора структуры аппаратно-программных средств в составе автоматизированных информационных комплексов.

МОДЕЛЬ ДАННЫХ В ФОРМАТЕ ДЕФИНИЦИЙ

Начнем описание понятия «модели данных» с его определения или дефиниции.

Дефиниция[3] – это уникальное логическое и языковое явление, в котором язык (как способ описания явления) и мышление (логика построения самого описания) взаимодействуют, совершенствуя друг друга [10]. Дефиницию можно назвать своеобразным «информационным каналом», связывающим имеющиеся знания с вновь полученными, что помогает специалистам различных ПрО понимать друг друга, обеспечивая доступность когнитивной деятельности человека.

Структура МД представляет собой разметку содержимого информационной модели, характерной для конкретного типа репозитория (хранилища), протокола обмена, платформы и т.д. (в соответствии с технологиями представления, организации, хранения и управления данными[4]). Иными словами, это некоторая «абстрактная машина доступа к данным, с которой взаимодействуют» потребители информации [11] и которая необходима для создания баз данных. Чаще всего это иерархический список объектов и описание связей между ними. Включение ресурсов в группы, групп в задачи, а задач – в задания называется установлением связей между объектами, составляющими суть понятия «моделей данных», представленных в табл. 1. Основной смысл большинства определений – в представлении данных в виде структуры, для которой важны связи, обеспечивающие логику применения данных, например, при формировании таблиц «температур» окружающей среды. Когда, например, данные «температура» интегрируются с данными «шкалы» – записями единиц измерения (по Цельсию, Фаренгейту, Кельвину), то есть с метаданными, возникает эффект семантической интеграции данных, что удобно для пользователей базы данных, но создает сложности при организации хранения данных.

Таким образом, МД должна характеризоваться:

- структурой данных (таблицы, списки, деревья);
- совокупностью возможных операций с данными (поиск, обновление, сбор и т.д.);
- набором правил (отношений) для выстраивания соединений между данными.

[2] Сайт PromoPuit.ru [Электронный ресурс]. URL: https://promopult.ru/subscribe.html?id=217 (дата обращения 13.06.2024).

[3] Дефиниция (от лат.definitio) – определение, точное указание, раскрывающее содержание (смысл) понятия или концепции.

[4] ГОСТ Р 56174–2014. Информационные технологии. Архитектура служб открытой грид-среды. Термины и определения. – М.: Стандартинформ, 2014. (п. 3.1.42).

Таблица 1

Дефиниции понятия «модель данных»

№ п/п	МОДЕЛЬ ДАННЫХ (МД) – ЭТО…	ИСТОЧНИКИ
1	Абстрактное, самодостаточное, логическое определение объектов, операторов и прочих элементов, в совокупности составляющих абстрактную машину доступа к данным, с которой взаимодействует пользователь. Эти объекты позволяют моделировать структуру данных, а операторы – поведение данных	Дейт К. Дж. [11]
2	Совокупность правил порождения структур данных в базе данных, операций над ними, а также ограничений целостности, определяющих допустимые связи и значения данных, последовательность их изменения	ГОСТ 20886–85; (п. 58, с. 5)
3	Набор конструктивов, обеспечивающих определение, структуру и формат данных; этот набор может быть физическим или абстрактным, в зависимости от выбора регистрирующей среды	ГОСТ Р 55345–2012/ ISO/TS 18876-2:2003; (п. 3.1.6)

Продолжение табл. 1

№ п/п	МОДЕЛЬ ДАННЫХ (МД) – ЭТО…	ИСТОЧНИКИ
4	Схема данных, структурированная в базе данных в соответствии с формальными описаниями в информационной системе и требованиями используемой системы управления базой данных	ГОСТ Р ИСО/МЭК 20546–2021; (п. 3.1.5)
5	МД – разметка содержимого информационной модели по форме, смоделированной для конкретного типа репозитория, протокола, платформы и т. д., и представленная информационной моделью в соответствии со спецификацией набора механизмов для представления, организации, хранения и управления данными	ГОСТ Р 56174–2014; (п. 3.1.42)
6	МД – графическое и/или лексическое представление данных, устанавливающее их свойства, структуры и взаимосвязи	ГОСТ Р ИСО/МЭК 19778-1–2011; (п. 3.1.7)
7	Концептуальная модель данных – МД, которая представляет абстрактную точку зрения на реальный мир	ГОСТ Р ИСО/МЭК 19778-1–2011; (п. 3.2.5)

С другой стороны, МД можно назвать агрегатом данных, характеризуемым векторной или иерархической структурой [12].

К наиболее распространенным моделям данных можно отнести следующие модели [11, 12]:

- иерархические модели данных (ИМД), для которых характерна древовидная наглядная структура, объединяющая объекты различных уровней;
- реляционные модели данных (РМД), представляющие собой логический тип моделей;
- объектно-реляционные модели данных (ОРМД), построенные на основе принципов объектно-ориентированного программирования;
- сетевые модели данных (СМД);
- многомерные модели данных (ММД).

Сравнительный анализ перечисленных моделей данных представим в виде табл. 2.

Наиболее существенными параметрами БД являются следующие:

- быстродействие;
- особенности обновления данных; независимость данных;
- возможность работы в многопользовательском режиме выдачи данных; безопасность данных и БД в целом;
- стандартизация построения и эксплуатации БД;

Таблица 2

Сравнительный анализ моделей данных

ОСНОВНЫЕ ХАРАКТЕРИСТИКИ	ИМД	РМД	СМД	ММД
Объем требуемой памяти	Минимальный	Большой	Средний	Небольшой
Скорость выполнения операций соединения	Средняя	Низкая	Высокая	Высокая
Простота организации	Да	Да	Нет	Нет
Универсальность	Доступ через концевой объект	Да	Да	Да
Независимость данных	Да	Да	Да	Да
Принципы доступа к данным	Только навигация	Логический	Только навигация	Только навигация
Теоретическое обоснование методов, отношений	Нет	Да	Да	Да
Особенности структуры	Простая	Гибкая	Сложная	Сложная
Особенности программирования	Не сложное	Сложная	Не сложное	Не сложное
Число элементов	Достаточное	Избыточное	Достаточное	Достаточное

– адекватность отображения данных, соответствующих заданной предметной области.

Основные элементы модели данных представлены на рис. 1.

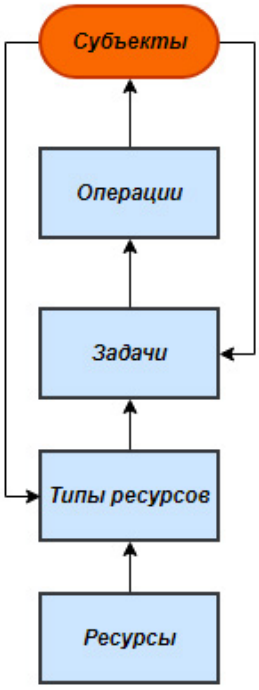

Рис. 1. Модель руководства организацией (командой)

Модель данных состоит из двух частей, одна часть относится к замкнутой программной среде, другая – к контролю целостности. Пользователи (субъекты) формируют задания, состоящие из подзадач (процессов, операций), выбирают ресурсы в своей ПрО (вычислительные, вспомогательные и др.). Модель может быть задействована и фрагментарно в зависимости от решаемой задачи [13].

ПОКАЗАТЕЛИ КАЧЕСТВА БОЛЬШИХ ДАННЫХ

Аналитика больших данных постоянно расширяется по мере развития методов машинного обучения, интеграции с облачными технологиями, активным внедрением инструментов цифровых двойников, повышением требований к безопасным технологиям и другим тенденциям в области искусственного интеллекта[5].

Популярность больших данных обычно рассматривается как дополнение к процессам управления в различных отраслях. Анализ больших данных может улучшить операционные и стратегические возможности предприятия, такие как бизнес-анализ, управление цепочками поставок и промышленные процессы, улучшить принятие решений на различных этапах производства [14].

Качество данных можно определить с точки зрения их свойств, называемых показателями качества данных. Шкалы метрик качества данных в большинстве случаев лежит в интервале или в соответствующих процентных значениях. Многие из показателей качества данных, рассмотренные в части 1 настоящей работы [1], остаются действительными и для BD. Однако, с появлением больших данных, были поставлены под сомнение некоторые понятия качества данных, такие как применимость существующих показателей, эффективность инструментов оценки и точность измерений. Таким образом, приведем примеры формализации показателей качества в контексте BD (см. табл. 3), и покажем, какие характеристики больших данных влияют на эти показатели. В [1] приведены 12 показателей качества применительно к BD, основываясь на [15]. Это полнота, своевременность, изменчивость, уникальность, согласованность, соответствие (валидность), простота манипулирования, релевантность, читабельность, безопасность, доступность и целостность.

Под безопасностью данных будем понимать степень ограничения доступа (защиту от несанкционированного доступа[6]) к данным. В связи с ростом масштабных нарушений конфиденциальности и атак на систему безопасности обеспечение конфиденциальности и безопасности данных стало одним из приоритетных направлений в обеспечении качества BD. Здесь мы приведем лишь один из возможных подходов к оценке безопасности.

Для этого в [15] предлагается методом опроса присвоить «веса» составляющим признакам безопасности для конкретной, например, БД. При этом вклад каждого признака будем считать равноценным (одинаковым).

Вопросы анкеты для оценки уровня безопасности данных следующие:

1. Существует ли политика безопасности, ограничивающая использование данных?
2. Используются ли протоколы безопасности для передачи данных?
3. Существуют ли меры по обнаружению угроз?
4. Зашифрованы ли данные надлежащим образом?
5. Имеется ли документация по безопасности, сопровождающая данные?

Выражение для безопасности данных примет вид

$$\text{Безопасность } (\%) = \sum_{i=1}^{5} b_i \, c_i, \qquad (1)$$

где b_i – весовые множители, для случая равнозначности признаков все b_i=20% $i = \overline{1,5}$; c_i; – оценки выбранных пяти признаков по шкале от 0 до 1.

[5] Комлев М. 7 трендов в аналитике больших данных // Tadviser [сайт]. – URL: https://www.tadviser.ru (дата обращения: 13.06.2024).

[6] См. ГОСТ 33707–2016 Информационные технологии. Словарь. – М.: Стандартинформ, 2016. (Введ.09-01-2017). (п. 4.90).

Таблица 3

Основные показатели (меры) качества больших данных

№ п/п	СОДЕРЖАНИЕ / ПРЕИМУЩЕСТВА ПРИЗНАКА	ПОКАЗАТЕЛИ КАЧЕСТВА БОЛЬШИХ ДАННЫХ
1	Степень, в которой данные достаточно полны и содержат необходимую информацию	$\text{Полнота}\ (\%) = \frac{\text{Количество непустых значений}}{\text{Общее число значений}} \times 100$
2	Уникальность данных можно определить как соотношение неповторяющихся значений, т.к. дублированные записи искажают аналитические результаты.	$\text{Уникальность}\ (\%) = \frac{\text{Количество уникальных строк}}{\text{Всего строк}} \times 100$
3	Данные, представленные в одной и той же структуре и соответствующие схемам и стандартам данных	$\text{Согласованность}\ (\%) = \frac{\text{Количество значений с согласованными типами}}{\text{Общие значения}} \times 100$
4	Степень соответствия данных правилам и ограничениям своей среды (шаблонам полей с определенным синтаксисом)	$\text{Своевременность}\ (\%) = \frac{\text{Текущая дата – Дата последнего изменения}}{\text{Текущая дата – Дата создания}} \times 100$
5	Может быть определена как задержка между текущей датой и датой последнего изменения данных	$\text{Своевременность}\ (\%) = \frac{\text{Текущая дата – Дата последнего изменения}}{\text{Текущая дата – Дата создания}} \times 100$
6	Как долго данные могут храниться и считаться достоверными, т.е. это задержка между датами хранения и изменения данных	$\text{Изменчивость}\ (\%) = \frac{\text{Дата создания – Дата изменения}}{\text{Текущая дата – Дата создания}} \times 100$
7	Читабельность – способность обрабатывать и извлекать информацию, содержащуюся в данных, включая аудио и видео форматы	$\text{Читабельность}\ (\%) = \frac{\text{Количество обработанных значений без ошибок}}{\text{Общие значения}} \times 100$
8	Простота манипулирования (ПрМ) – степень использования с минимальными усилиями, включая время на подготовку (очистку, интеграцию данных, сокращение)	$\text{ПрМ}\ (\%) = \frac{\text{Количество различий между исходной и очищенной таблицей}}{\text{Общие данные}} \times 100$
9	Данные с большим количеством обращений считаются наиболее релевантными	$\text{Релевантность}\ (\%) = \frac{\text{Количество доступов к полю данных}}{\text{Полный доступ к таблице данных}} \times 100$
10	Безопасность данных, как степень надлежащего ограничения доступа к данным	$\text{Безопасность}\ (\%) = \sum_{i=1}^{5} b_i\, c_i.$
11	Данные доступны и их легко получить	$\text{Доступность}\ (\%) = \frac{\text{Количество доступных значений}}{\text{Общие значения}} \times 100$
12	Целостность данных означает точность и достоверность данных на протяжении их жизненного цикла	$\text{Целостность}\ (\%) = \frac{\text{Количество различий между исходными и обработанными данными}}{\text{Общие значения}} \times 100$

Целостность данных означает точность и достоверность данных на протяжении их жизненного цикла. В среде BD данные перед использованием проходят последовательные этапы преобразований (переработки). Поэтому важно гарантировать, что значения данных не были изменены и что достоверность данных при этом сохраняется. Измерение целостности состоит из сравнения значений данных до и после обработки данных (см. табл. 3). Таким образом, мы определяем целостность данных как отношение различий между исходными и обработанными значениями данных к общим значениям.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ БОЛЬШИХ ДАННЫХ

Основной рабочий механизм концептуальной основы интеллектуального анализа может моделировать, классифицировать и агрегировать данные, обнаруживать корреляцию между данными [14], что позволяет использовать эти результаты в подсистемах поддержки принятия решений [8] при управлении сложными технологическими процессами, включая технологии предиктивного управления, основанного на управлении данными в сочетании с методами анализа данных.

Механизм анализа больших данных в сложной организационно-технической производственной структуре включает следующие этапы [1, 14], представленные на рис. 2:

- сбор операционной системой разнородных (в общем случае) данных;
- хранение и пополнение данных в БД;
- очистка данных;
- интеграция данных;
- анализ данных;
- выявление практически полезных знаний средствами Data Mining;
- визуализация результатов анализа.

Для исследования системы ИАД (здесь и далее будем подразумевать под данными именно большие данные) воспользуемся концепцией многоагентного представления технологий объектно-ориентированного программирования [3, 4]. В данном подходе выполнение задач на каждом из этапов поручается некоторому агенту (программе), обеспечивающему одновременно и коммуникативные функции взаимодействия между агентами и внешней средой. Любой из агентов обладает свойствами автономности, реактивности, целеполагания, коммуникативности, обучаемости.

Для представленных на рис. 2 этапов функционирования системы ИАД введем обозначения для соответствующих агентов:

- $A_{СД}$ агент сбора данных, работа которого состоит в формировании первичного набора данных на основании предварительного экспертного оценивания потоков данных (источников данных), необходимых для выполнения целевых задач;
- $A_{БД}$ агент базы данных осуществляет распределение и консолидацию данных в БД для удобства дальнейшего анализа;
- $A_{ОД}$ агент очистки данных, обеспечивающий предварительную обработку данных (выявление и устранение пропусков данных (сглаживание)), исключение дубликатов, сжатие (применительно к медленно меняющимся процессам) и др.;

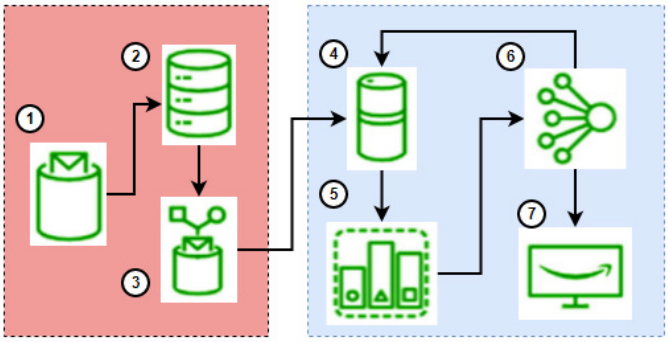

Рис. 2. Структура системы интеллектуального анализа больших данных

- $A_{ИД}$ агент интеграции данных, задача которого – обеспечить совместимость между различными технологиями, возможность слияния данных применительно, как для одной и той же сущности реального мира (ПрО), но полученных из разных источников данных [16], так и для других задач (по запросам);
- $A_{АД}$ агент анализа данных Анализ больших данных может динамически воспринимать и отслеживать процесс производства продукции, прогнозировать рыночный спрос и обеспечивать реконструкцию и оптимизацию производственного процесса (применительно к системам ИАД в производственном секторе);
- A_{DM} агент ИАД (средствами Data Mining), в основе которых методы обнаружения знаний в составе задач классификации, кластеризации, ассоциации, методов регрессионного анализа и ряда других;
- $A_{ВД}$ агент визуализации данных, предназначенный для наглядного представления результатов анализа в заданном формате для авторской презентации и поддержки полученных выводов.

Представим процессы взаимодействия агентов в системе ИАД в виде многоагентной системы (MAS) [4]:

$$MAS = (A, R, St_{ORG}), \quad (2)$$

где $A=\{A_i\}$ – множество разнотипных агентов, $i = \overline{1,N}$, а N – общее число типов агентов, основные из которых получили буквенные индексы для понимания сущности выполняемых ими задач; $St_{ORG} = \{st_j\}, j = \overline{1,S}$, – множество организационно-информационных структур, причем любая структура соответствует текущей выполняемой аналитической задаче, характеризуемой выбранным набором данных (или набором интегрируемых данных), методом анализа и другими характеристиками;

R – семейство базовых отношений между агентами:

$$R = R_1 \cup R_2 \cup R_3, \quad (3)$$

где $R_1 = \{1 \to 2; 2 \to 3; \dots; 6 \to 7\}$ – множество последовательных (типовых) переходов между этапами; $R_2 = \{6 \to 4; \dots; 7 \to 4\}$ –

множество функциональных и управляющих запросов для оптимизации текущего решения; R_3– множество отношений между агентом координатором из общего состава типов агентов N и любой группы агентов: a) подмножества агентов {1; 2; 3}, осуществляющих формирование и предобработку данных; b) подмножества агентов {3; 4;…; 7}, отвечающих за этапы ИАД.

Процессы взаимодействия агентов из множества представим кортежем:

$$Int = \langle A, T, Pr \rangle , \quad (4)$$

где T – множество типов агентов, т.е. T=$\{A_{СД}, A_{БД}, A_{ОД}, A_{ИД}, A_{АД}, A_{DM}, A_{ВД}\}$ – агенты – исполнители задач, указанные выше; Pr – сценарии или программы взаимодействия между агентами, причем

$$Pr = (\mathrm{Com}, \pi),$$

где Com – множество коммуникативных действий между агентами (СД→ДБ; ДБ→ОД; и др.); π – протоколы типовых действий (запись данных, копирование данных, передача данных, сжатие данных и т. д.), свойственных уровню взаимодействия (отношений) из (3).

Особенностью представления MAS в виде (2) – (4) является то, выстраиваемая структура отношений и соответствующих коммуникаций каждый раз определяется особенностями решаемой задачи и степенью достижения цели.

Данный подход позволяет расширять функционал модели, за счет ввода новых типов агентов, например, в ходе реализации новых методов оценивания данных, а также расширения объема привлекаемых BD.

ЗАКЛЮЧЕНИЕ

Таким образом, рассмотрен класс моделей данных для оценки качества больших данных в практических приложениях для задач поддержки принятия решений при формировании баз данных и организации комплексирования данных.

Использование агентного моделирования позволяет на концептуальном уровне проанализировать возможности синергетического эффекта больших данных по выявлению дополнительных знаний для повышения эффективности поддержки управляющих факторов в ходе принятия решений в сложных организационно-технических системах.

Дальнейшими направлениями исследований, на наш взгляд, являются:

- оценка информационно-коммуникационного аспекта данных, как способа передачи информации, и роль качества данных в этом процессе;
- анализ больших данных как продукции при использовании (технико-социальный аспект), когда следует применять не только рассмотренные показатели качества данных, но и вводить новые, отражающие специфику ряда технических задач (функциональные особенности – готовности, эффективности, полезности, устойчивости данных).

Заметим, что большинство показателей качества данных построены для идеальной среды, надежность и безопасность данных рассматривается лишь на входе в БД (в виде систем паролей, уровней конфиденциальности и др.). «За кадром» остаются структурные особенности данных и их влияние, например, на безопасность и защиту информации, и безопасность информационных технологий в целом.

Списки использованных источников и литературы

1. 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.
2. Когаловский М.Р. Перспективные технологии информационных систем. – М.: ДМК Пресс, 2018. – 288 с.
3. Бурый А.С., Ловцов Д.А. Информационные структуры умного города на основе киберфизических систем // Правовая информатика. 2022. № 4. С. 15–26.
4. Аронов И.З., Бурый А.С., Рыбакова А.М. Умная экономика замкнутого цикла: основа цифровых стратегий производственных компаний. Часть 2. Циркулярные бизнес-модели // Информационно-экономические аспекты стандартизации и технического регулирования. 2022. № 5 (69). С. 17–26.
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. Китчин Р. Управляемый данными сетевой урбанизм // Шаги / Steps. 2017. Т. 3, № 2. С. 98–116.
7. Бурый А.С., Шевкунов М.А. Суррогатное моделирование распределенных информационных систем по большим данным // Информационно-экономические аспекты стандартизации и технического регулирования. 2019. № 5(51). С. 43–50.

8. Особенности применения принципов интеллектуального анализа данных в корпоративных информационных системах / А.И. Бачурин, А.В. Мельников, А.А. Распопов, Д.Т. Шкубулиани // Информационно-экономические аспекты стандартизации и технического регулирования. 2021. № 4 (62). С. 39–44.
9. Питкевич П.И. Разработка структуры WEB-системы обработки больших данных // Universum: технические науки. 2021. № 12–1 (93). С. 75–78.
10. Гришечкина Г.Ю. Виды дефиниций терминов в научно-популярном тексте // Ученые записки Орловского государственного университета. Серия: Гуманитарные и социальные науки. 2010. № 1 (35). С. 120–127.
11. Дейт К.Дж. Введение в системы баз данных. Пер. с англ. – М.: ООО «И.Д. Вильямс», 2016. – 1328 с.
12. Бурый А.С., Морин Е.В. Модельно-алгоритмические структуры оценки качества программных изделий. – М.: «Горячая линия-Телеком», 2019. – 160 с.
13. Гохович В.А., Воробьев Э.А.,. Бурмакин А.О. Структура модели данных: изолированная программная среда и механизм контроля целостности // Синергия Наук. 2019. № 38. С. 190–197.
14. Li C., Chen Y., Shang Y. A review of industrial big data for decision making in intelligent manufacturing // Engineering Science and Technology, an International Journal. 2022. Vol. 29. P. 101021.
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. № 6 (4). С. 153.
16. 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 2. Data models

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, network technologies, cloud services and a number of others. Continuing the research started in Part 1, the types of data models and their role in Big Data analysis are clarified.

The purpose of the study of the second part of the work is the analysis and applicability of data models in practical applications for modeling decision support processes and an informed choice of data structure for the tasks of their integration and intelligent analysis. For the task of intellectual analysis of Big Data based on a multi-agent approach, the main elements, a system of relations in the formation of stages of data processing in order to discover new knowledge, are identified.

Keywords: Big Data, data model, Big Data quality indicators, Data Mining, data integration.

For citation: Buryi A.S., Pogodin I.M. Assessment the Quality of Big Data. Part 2. Data Models. Information and Economic Aspects of Standardization and Technical Regulation. 2024; 4 (79): 24–32. (In Russ.).

References

1. Buryi A.S., Pogodin I.M. Ocenka kachestva bol'shih dannyh. Part 1. Osnovnye ponyatiya i metriki. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya. 2024, no. 3 (78), pp. 49–58. (In Russ.).
2. Kogalovskij M.R. Perspektivnye tekhnologii informacionnyh sistem. Moscow: DMK Press Publ., 2018. 288 p. (In Russ.).
3. Buryi A.S., Lovtsov D.A. Informacionnye struktury umnogo goroda na osnove kiberfizicheskih sistem. Pravovaya informatika. 2022, no. 4, pp. 15–26. https://doi.org/10.21681/1994-1404-2022-4-15-26 (In Russ.).

4. Aronov I.Z., Buryi A.S., Rybakova A.M. Umnaya ekonomika zamknutogo cikla: osnova cifrovyh strategij proizvodstvennyh kompanij. CHast' 2. Cirkulyarnye biznes-modeli. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya. 2022, no. 5 (69), pp. 17–26. (In Russ.).
5. Buryi A.S., Slepyntseva L.I. Cifrovizaciya kontenta dokumentov po standartizacii. Part 1. Sostoyanie i sovremennye tendencii. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya. 2021, no. 1 (59), pp. 105–113. (In Russ.).
6. Kitchin R. Upravlyaemyj dannymi setevoj urbanizm. Shagi / Steps. 2017, vol. 3, no. 2, pp. 98–116.
7. Buryi A.S., Shevkunov M.A. Surrogatnoe modelirovanie raspredelennyh informacionnyh sistem po bol'shim dannym. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya, 2019, no. 5(51), pp. 43–50. (In Russ.).
8. Bachurin A.I., Mel'nikov A.V., Raspopov A.A., Shkubuliani D.T. Osobennosti primeneniya principov intellektual'nogo analiza dannyh v korporativnyh informacionnyh sistemah. Informacionno-ekonomicheskie aspekty standartizacii i tekhnicheskogo regulirovaniya, 2021, no. 4 (62), pp. 39–44. (In Russ.).
9. Pitkevich P.I. Razrabotka struktury WEB-sistemy obrabotki bol'shih dannyh. Universum: tekhnicheskie nauki, 2021, no. 12–1 (93), pp. 75–78. (In Russ.).
10. Grishechkina G.Y. Vidy definicij terminov v nauchno-populyarnom tekste. Uchenye zapiski Orlovskogo gosudarstvennogo universiteta. Seriya: Gumanitarnye i social'nye nauki, 2010, no. 1 (35), pp. 120–127. (In Russ.).
11. Date C.J. Vvedenie v sistemy baz dannyh – Introduction to Database Systems. Moscow: OOO «I.D. Vil'yams» Publ., 2016. 1328 p. (In Russ.).
12. Buryi A.S., Morin E.V. Model'no-algoritmicheskie struktury ocenki kachestva programmnyh izdelij. Moscow: «Goryachaya liniya-Telekom», 2019. 160 p. (In Russ.).
13. Gohovich V.A., Vorob'ev E.A., Burmakin A.O. Struktura modeli dannyh: izolirovannaya programmnaya sreda i mekhanizm kontrolya celostnosti. Sinergiya Nauk, 2019, no. 38, pp. 190–197. (In Russ.).
14. Li C., Chen Y., Shang Y. A review of industrial big data for decision making in intelligent manufacturing. Engineering Science and Technology, an International Journal, 2022, vol. 29. Art. 101021.
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). Art. 153. https:// doi.org/10.3390/bdcc6040153
16. Vovchenko A.E., Kalinichenko L.A., Kovalev D.Y. 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.