УДК 004.6 ***DOI: 10.46960/1816-210X_2021_2_16***

КОНЦЕПТУАЛЬНАЯ МОДЕЛЬ БОЛЬШИХ ДАННЫХ

В.Ю. Карпычев
ORCID: **0000-0001-8527-2600** e-mail: **kavlyr@yandex.ru**
Нижегородский государственный технический университет им. Р. Е. Алексеева
Нижний Новгород, Россия

Ю.П. Шальнова
ORCID: **0000-0002-1953-3734** e-mail: **julia.shalnova@gmail.com**
ПАО «Сбербанк»
Нижний Новгород, Россия

Предложена концептуальная модель больших данных – абстрактная и обобщенная, но позволяющая понимать сущность моделируемого объекта и выступать основой разработки частных моделей больших данных для различных предметных областей. Назначением концептуальной модели является проведение технических и экономических исследований, направленных на повышение эффективности (монетизацию) и снижение рисков внедрения технологий больших данных в условиях нехватки информации, недоопределенных и некорректных прикладных задач. Сформулирована задача обработки больших данных. Показаны особенности и трудности ее постановки. Предложен набор параметров, составляющих модель больших данных, и критерии его формирования. Отмечен пилотный и итеративный характер разработанной модели.

Ключевые слова: большие данные, концептуальная модель, характеристики больших данных, экономическая эффективность.

ДЛЯ ЦИТИРОВАНИЯ: Карпычев, В.Ю. Концептуальная модель больших данных / В.Ю. Карпычев, Ю.П. Шальнова // Труды НГТУ им. Р.Е. Алексеева. 2021. № 2. С. 16-23. DOI: 10.46960/1816-210X_2021_2_16

BIG DATA CONCEPTUAL MODEL

V.Yu. Karpychev
ORCID: **0000-0001-8527-2600** e-mail: **kavlyr@yandex.ru**
Nizhny Novgorod state technical university n.a. R.E. Alekseev
Nizhny Novgorod, Russia

Yu.P. SHal'nova
ORCID: **0000-0002-1953-3734** e-mail: **julia.shalnova@gmail.com**
PAO «Sberbank»
Nizhny Novgorod, Russia

Abstract. Big data conceptual model is proposed; it is abstract and generalized, but it allows us to understand the essences of the modeled object, and to serve as the basis for development of big data models for various subject areas. Purpose of the conceptual model is to conduct technical and economic research focused on improving the efficiency (monetization), and reducing the risks of implementing the big data technologies in the context of lack of information, undetermined and incorrect application of tasks. Big data processing problem is formulated. Features and difficulties of its formulation are shown. A set of parameters that make up the big data model, and criteria for its formation are proposed. Pilot and iterative nature of the developed model is noted.

Key words: big data, conceptual model, energy of neutrons, big data features, economic efficiency.

FOR CITATION: Karpychev V.Yu., SHal'nova Yu.P. Big date conceptual model. Transactions of NNSTU n.a. R.E. Alekseev. 2021. № 2. P. 16-23. DOI: 10.46960/1816-210X_2021_2_16

Введение

Понятие «большие данные» исследовано во множестве научных работ. В качестве рабочего в данной статье будем использовать определение консалтинговой компании *Gartner*: «большие данные – это большой объем, высокая скорость и/или большое разнообразие информационных активов, которые требуют *экономически эффективных* (курсив наш) инновационных форм обработки информации, позволяющих улучшить понимание, принятие решений и автоматизацию процессов» [1]. Большие данные используются в различных задачах [2], в том числе, для добычи (*mining*) скрытой в больших данных информации (*hidden pattern*). Важно заметить, что существующие и скрытые в больших данных закономерности, взаимосвязи и тенденции изначально не предназначаются для выявления и использования. Укрупненно обработка больших данных включает этапы сбора, предподготовки, анализа и интерпретации данных [3,4]. В силу различных причин, не рассматриваемых в настоящей статье, внедрение в деятельность организации технологии больших данных является весьма высокорисковой и затратной инновацией. При этом этапы сбора и предподготовки данных не только в значительной мере определяют успешность проекта, но и вносят значительный вклад (до 70 % [5]) в затраты на его осуществление, хотя предназначение проекта реализуется на стадии анализа данных [6].

Серьезных научных исследований, направленных на повышение эффективности начальных стадий проектов больших данных, в отечественной специальной литературе не представлено. Существующие примеры внедрения технологий больших данных носят исключительно эксклюзивный (абсолютная индивидуальность конкретных решений) и проприетарный характер. Этот фактор накладывает ограничения на тиражирование внедренных проектов. Перечисленные детерминанты обуславливают актуальность исследований стадий постановки задачи и предподготовки больших данных, в том числе, в части обеспечения эффективности технологии и ее доступности заказчикам. При этом начальные этапы проектирования часто выполняются в условиях неопределенности требований к системе больших данных, что исключает использование аналитических, алгоритмических или качественных решений. Подобного рода исследования можно провести моделированием функционала и структурно-параметрических характеристик систем больших данных. Однако в силу рассмотренных выше особенностей, обобщенные модели больших данных до настоящего времени не созданы.

Таким образом, актуализируется задача разработки концептуальной модели больших данных, которая, с одной стороны, является абстрактной и достаточно обобщенной: отсутствуют особенности и детали конкретных решений по большим данным. С другой стороны, концептуальная модель должна быть основой для разработки частных моделей больших данных для различных предметных областей. Данное требование означает, что целью построения концептуальной модели больших данных в рассматриваемом контексте является использование ее для анализа и синтеза экономических, а также технических и управленческих решений.

Цели обработки больших данных

В работе [7] задача создания информационных систем формализована простой записью (1):

$$G = F(D). \qquad (1)$$

Применительно к системе больших данных: G – результат (цель) обработки множества входных данных D; F – функционал системы больших данных.

Обычно исходные данные D и требуемый результат G задаются предметным специалистом. Разработчик информационной системы должен определить и материализовать в аппаратно-программных решениях ее функционал. Постановка задачи, как мы отмечали, явля-

ется очень важным этапом, так как корректность и адекватность ее формулировки определяет успешность проекта больших данных, а ошибки и неточности ведут, в лучшем случае, к экономическим потерям. Для постановки задачи обработки больших данных, как правило, должны существовать осознанные потребность или необходимость в получении результата *G*. Данные есть прямое следствие выбранной цели. Невозможно обрабатывать большие данные до того, как будет определена цель.

Однако при разработке систем больших данных типичной является ситуация, в которой ни множество данных *D*, ни цель (назначение) *G* системы четко не определены. Более того, возможна ситуация, когда и потребность (цель) не определена (существует, но не осознаваема). В этом случае используются специальные методы анализа больших данных.

Формирование множества больших данных

Постановку задачи (1) следует начинать с формулирования цели *G* обработки больших данных. Наличие цели *G* позволяет сформировать множество больших данных *D*. Эту процедуру можно формализовать записью (2):

$$G \to D = \{d_1,...d_m\}, \quad (2)$$

где d_i принадлежит множеству предметно-ориентированных данных *D*.

Выполнение процедуры (2) проводится предметными специалистами с учетом семантической принадлежности (релевантности) элемента d_i к предметной области и степени его влияния на достижение цели *G* (важности). Цели также определяют источники и методы анализа данных.

Решение задачи (2) сопряжено с трудностями в части определения:

- важности и источников некоторых данных;
- объема, релевантности и качества скрытой в исходном множестве *D* информации.

Существует также не рассматриваемая в настоящей статье проблема определения сигнальных признаков элемента d_i. В данном случае *сигнал* – это формальный признак, который косвенно характеризует потенциальную полезность скрытой информации для целей обработки больших данных. Можно согласиться, что «найти данные для анализа – это отчасти наука, отчасти исследовательская работа и отчасти предположения» [8]. Эта особенность постановки задачи (1) во многом определяет успешность проекта больших данных.

Семантика и синтаксис больших данных

В общем случае содержание, смысл и назначение фрагментов объективного мира находят отражение в некоторых воспринимаемых человеком материальных сущностях: речи, тексте (символы), изображениях, аудио- и видеофайлах. При этом предполагается, что семантика информации доступна для восприятия и обработки человеком непосредственно или с использованием специального инструментария (в настоящей статье – автоматизированные информационные технологии). При таком подходе работа с большими данными предполагает экспертные: постановку цели *G*, формирование множества *D* и определение методов обработки. Непосредственную обработку исходных данных и обнаружение скрытой в них информации осуществляет ЭВМ. Поэтому определяющее значение для создания системы больших данных имеет понятие «семантическая единица» [9] (СЕ), под которой мы понимаем неделимую, однозначно определяемую смысловую сущность d_i независимо от ее организации, формы представления и технической реализации. Другими словами, мы отождествляем понятия семантической единицы и элемента множества данных d_i, обеспечивая тем самым человеко-машинный интерфейс.

При таком подходе под большими данными *D* можно понимать множество семантических единиц D={di}, детерминированных G, т.е., отвечающих условию (2). Таким образом, задача (2) может быть интерпретирована как спецификация множества СЕ. Как было отме-

чено ранее, концепция больших данных предполагает автоматизированную обработку доступной для человеческого восприятия семантики. Теоретически эта задача должна решаться путем формализации семантической информации (СЕ). Однако это очень сложная задача. «При формализации изучаемым объектам, их свойствам и отношениям ставятся в соответствие некоторые устойчивые, хорошо обозримые и отождествимые материальные конструкции, дающие возможность выявить и зафиксировать существенные стороны объектов» [10]. Поэтому далее под семантикой данных будем понимать формальное описание смысла данных. Степень формализации семантики данных можно использовать для отнесения их к категориям структурированных, полуструктурированных и неструктурированных данных [6].

Структурирование – это выделение в информации важных для достижения цели *G* СЕ и фиксация связей между ними. Для подготовки и анализа структурированных данных существует хорошо развитой математический аппарат. Для идентификации СЕ используются:

- *имя* – условное обозначение;
- *значение* – параметр, характеризующий свойства сущности, обозначаемой СЕ;
- *тип* – множество значений СЕ, сгруппированных по определенным признакам и в соответствии с перечнем допустимых преобразований.

В неструктурированных данных СЕ и связи не идентифицированы. Существует множество видов (форматов) неструктурированных данных: текстовые, числовые, графическая информация, аудио и видеофайлы. Поэтому для автоматизированной обработки неструктурированных данных используются схема описания предметной области и правила определения релевантных ей данных (онтология). Схема должна содержать СЕ и связи между ними. Семантические единицы определяются набором специальных правил. Связи также должны быть поименованы и могут содержать различные атрибуты, описывающие, например, способ связи. Следует отметить, что одни и те же данные могут рассматриваться как структурированные или неструктурированные в зависимости от предметной постановки задачи. Так, результаты измерения температуры за период времени можно представить в виде таблицы, текста, чисел, графики, видеоряда и т.д.

Кроме семантической составляющей, в модели больших данных следует отразить синтаксическую адекватность, т.е., «формально-структурные характеристики информации, не затрагивающие ее смыслового содержания» [11]. Характеристики физически представляются конкретными форматами файлов, которые в силу исторических и технических причин практически исключают возможность инвариантной обработки в ЭВМ.

Интеграция данных

Анализу больших данных обычно предшествует их интеграция, которая в общем случае заключается в объединении данных разнородных источников. Разнородность проявляется в отличиях форматов данных и их схемах. Результат интеграции – единое представление полного и непротиворечивого множества данных, обеспечивающее их обработку [12]. Практическая интеграция подмножеств, структурированных и неструктурированных данных в настоящее время сопряжена с техническими трудностями. Поэтому анализ данных обычно проводят раздельно и, соответственно, получают скрытые данные (закономерности), присущие этим подмножествам. Однако концепция больших данных не допускает подобных упрощений: анализ структурированных и неструктурированных данных из различных источников должен быть проведен совместно. Для этого неструктурированные данные должны быть структурированы и приведены к единому формату. Интеграция данных возможна на синтаксическом, логическом и семантическом уровнях.

Синтаксическая модель интеграции предполагает приведение данных из различных источников к единому формату физического представления. Логический уровень интеграции обеспечивает оперирование данными различных источников (например, объектных баз данных, веб-сайтов и т.д.) в терминах единой глобальной схемы. На семантическом уровне ин-

теграции данные имеют семантически значимое единое представление в рамках онтологии предметной области (например, пересечение понятийных баз разных источников). При интеграции данных на синтаксическом и логическом уровнях семантические свойства данных не учитываются. Видимо, поэтому некоторые исследователи говорят о двух уровнях интеграции: синтаксическом и семантическом [13].

Примером интеграции структурированных и неструктурированных данных может быть задача «Аналитика телефонных переговоров» со следующими целями:

- выявление трендов телефонного трафика (структурированная информация);
- распознавание содержания отдельных коммуникаций (неструктурированная информация);
- выделение новой (скрытой) информации на множестве структурированных и неструктурированных данных.

В табл. 1 приведены наборы возможных семантических единиц, выделяемых на подмножествах, структурированных D^1 и неструктурируемых данных потока телефонных звонков D^2.

Таблица 1.

Интеграция структурированных и неструктурированных данных

Table 1.

Integration of structured and unstructured data

Аудиоаналитика (телефонные звонки, например IP-телефонии)		
	Стандартный статистический анализ на структурируемой информации, D^1	Семантический анализ на неструктурируемой информации (речи), D^2 Полнотекстовое распознавание речи (робот)
Семантические единицы, d_i	Количество, длительность, стоимость, входящие, исходящие, переадресация, нет дозвона, непринятые коммуникационные данные и др.	Речевые шаблоны (вариации конкретных слов), в том числе, тишина более 5 с, например, слова из словаря на 500 слов для оценки эмоций/трендов
Цель G	Интегрированные данные: взаимосвязь речевых шаблонов со структурированными данными	

В научной литературе описаны методы интеграции данных. Консолидация данных включает извлечение (extract) данных из источников, преобразование (transform) и загрузку (load) их в хранилище. При федерализации данные извлекают из разнородных источников при поступлении внешних запросов. Распространение данных предполагает их передачу (копирование) в оперативном режиме [14,15]. Параметр «метод интеграции данных» также может быть включен в концептуальную модель больших данных.

Иные параметры больших данных

Одним из предназначений предлагаемой концептуальной модели может быть исследование экономических характеристик технологии больших данных. Для этого модель целесообразно расширить рядом параметров. Кратко рассмотрим их.

Режим обработки данных. С технологической точки зрения, обработку данных кратко можно определить как любое преобразование данных при решении конкретной задачи, которое может происходить в различных режимах: пакетном, реального масштаба времени, разделения времени и др. [16]. Выбор режима определяется характеристиками источников данных, информационной системой обработки данных и решаемых в ней задач, включая задачу (1). При этом режим обработки данных детерминирует время, стоимость и другие важные «нетехнологические» характеристики системы обработки больших данных.

Темпоральные (временные) характеристики данных. В модели больших данных для некоторых задач следует включать темпоральные характеристики: момент, период, интервал [17]. Эти характеристики отражают как предметно, так и технологически значимые особенности систем больших данных, например:

- различные моменты создания данных и завершения их обработки (время заключения сделки и расчет продавца с покупателем при биржевой торговле);
- асинхронное время работы процессов обработки данных (в различных часовых поясах);
- различные временные интервалы работы с данными у параллельно выполняемых процессов;
- загрузка данных с временной задержкой, «задним числом» (в том числе пакетная передача данных);
- различные временные модели интегрируемых данных (дискретное и интервальное время) и др.

Эти временные особенности обработки данных прямо влияют на экономические характеристики систем больших данных.

Планируемый вид анализа данных. В отечественной и зарубежной специальной литературе отмечена вариативность анализа данных [18,19] (табл. 2). Вид анализа для конкретной задачи (1) следует определять на стадии ее постановки.

Таблица 2.

Виды анализа данных

Table 2.

Types of data analysis

Вид анализа	**Характеристика вида анализа (пример)**
Описательный (Descriptive)	Аналитика произошедших событий (статистический анализ телефонных коммуникаций)
Диагностический (Diagnostic)	Выявление зависимостей различных событий и факторов (повременной объем телефонного трафика, ограничение доступа к услуге связи в определенное время)
Прогнозный (Predictive)	Определение трендов на основе имеющихся данных, включая оценку вероятности возможных событий (отказ в обслуживании при определенных сочетаниях количества и длительности телефонных коммуникаций)
Предписывающий (Prescriptive)	Определение возможных сценариев поведения системы при изменении множества исходных данных и условий задачи (изменение телефонного трафика при изменении тарифной политики)

Технология *Data Mining*, предназначенная для нетривиального обнаружения в данных скрытой, ранее неизвестной, доступной для интерпретации и практически полезной информации [20], оперирует описательными и предсказательными классами моделей анализа, которые формализуют задачу (1). Методы *Data Mining*, соответствующие модели анализа, определяют требования к характеристикам исходных данных *D*, например, тип данных: численный, качественный и др. [4,21]. Это означает, что в концептуальную модель больших данных следует включить параметр «Предполагаемый метод анализа».

Принадлежность данных. Источники данных, как указывалось, детерминированы целью *G* и уникальны для каждой задачи (1). Поэтому при ее постановке всегда возникает вопрос об источниках в контексте обладания правами собственности на данные, хотя с точки зрения гражданского права данные (информация) не являются объектом правоотношений [22]. В нашей модели данные рассматриваются как экономический объект, который может быть описан такими характеристиками, как цена приобретения или стоимость хранения в зависимости от принадлежности данных. Эта экономико-правовая характеристика данных мо-

жет принимать значения: внутренние и внешние данные. *Внутренние данные* – это созданные или приобретенные данные, хранящиеся в организации, которые она может использовать по своему усмотрению.

Право собственности на *внешние данные* принадлежит третьим лицам (сторонним организациям). Доступ и использование внешних данных ограничены какими-либо условиями, например, по критерию возмездности можно выделить условно-бесплатные и коммерческие данные. В концептуальную модель больших данных также могут быть включены иные известные характеристики информации, например, достоверность, качество, толерантность данных и др. [18].

Заключение

Предложенная концептуальная модель больших данных включает набор параметров, влияющих на технические и экономические характеристики технологии больших данных. Предполагается проведение экономических исследований, направленных на повышение эффективности технологий больших данных.

Представленная версия модели является пилотной. При разработке конкретных (частных) решений модель может менять структуру, постепенно формируя типовую онтологию больших данных.

Библиографический список

1. Glossary Gartner. [Электронный ресурс]. URL: https://www.gartner.com/en/information-technology/glossary/big-data (дата обращения: 12.02.2020).
2. McKinsey Global Institute. Big data: The next frontier for innovation, competition and productivity. [Электронный ресурс]. URL: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation (дата обращения: 10.10.2020).
3. **García, S.** Big data preprocessing: methods and prospects / S. García, S. Ramírez-Gallego, J. Luengo et al. // Big Data Anal. 2016. 1, 9. [Электронный ресурс]. URL: https://doi.org/10.1186/s41044-016-0014-0 (дата обращения: 15.03.2020).
4. **Kantardzic, M.** Data mining: Concepts, Models, Methods, and Algorithms / M. Kantardzic // Wiley. Hoboken, 2020. – 661 с.
5. IBM. The biggest data challenges that you might not even know you have. [Электронный ресурс]. URL: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/ (дата обращения: 12.10.2020).
6. **Vozábal, M.** Tools and Methods for Big Data Analysis / M. Vozábal. – Pilsen: University of West Bohemia, 2016. – 134 с.
7. **Шальнова, Ю.П.** Монетизация больших данных: технико-экономический анализ драйверов роста и издержек / Ю.П. Шальнова // Экономика. Информатика. 2020. Т. 47. № 3. С. 491-500.
8. **Ohlhorst, F.** Big Data Analytics Turning Big Data into Big Money / F. Ohlhorst. – Wiley, 2013. – 176 с.
9. ГОСТ Р ИСО 15531-1-2008: Промышленные автоматизированные системы и интеграция. Данные по управлению промышленным производством. Часть 1. Общий обзор. – М.: Стандартинформ, 2009. – 20 с.
10. **Ивин, А.А.** Словарь по логике / А.А. Ивин, А.Л. Никифоров. – М.: Туманит, ВЛАДОС, 1997. – 384 с.
11. Информатика: концептуальные основы. – М.: Маросейка, 2008. – 464 с.
12. **Dong, X.L.** Big Data Integration / X.L. Dong, D. Srivastava. – Morgan&Claypool, 2015. – 178 с.
13. **Черняк, Л.** Интеграция данных: синтаксис и семантика // Открытые системы. СУБД. 2009. [Электронный ресурс] // Режим доступа: https://www.osp.ru/os/2009/10/11170978/ (дата обращения: 12.10.2020).
14. **White, C.** Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise / C. White // DM Review. 2005. № 11. С. 25-43.

15. **Кресов, А. А.** Принципы интеграции данных в сфере недропользования / А.А. Кресов, В.В. Уваров // Вестник кибернетики.2010. № 10. С. 83-89.
16. **Шепелев, К.В.** Анализ режимов автоматизированной обработки данных / К.В. Шепелев [и др.] // Промышленные АСУ и контроллеры. 2019. № 12. С. 48-53.
17. **Jensen, C.S.** Temporal Database Management / C.S. Jensen. **–** Aalborg University, 2000. – 1323 с.
18. **Anderson, C.** Creating a Data-Driven Organization / C. Anderson. – O'Reilly Media, 2015. – 302 с.
19. **Gerber, D.** 4 Types of Data Analytics / D. Gerber // Oracle AI and Data Science Blog. [Электронный ресурс]. URL: https://blogs.oracle.com/datascience/4-types-of-data-analytics (дата обращения: 15.09.2020).
20. **Frawley, W.** Knowledge Discovery in Databases: An Overview / W. Frawley, G. Piatetsky-Shapiro, C. Matheus // AI Magazine. 1992. С. 213-228.
21. **Барсегян, А.А.** Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: БХВ-Петербург, 2007. – 384 с.
22. **Карпычев, В.Ю.** Информация как объект гражданских правоотношений. В сб: Актуальные проблемы частного и публичного права / В.Ю. Карпычев, М.В. Карпычев // Всерос. научно-практ. конф. (Волгоград, 27 октября 2017 г.). Волгоград, ВАМВД России. – С. 73-76.

Дата поступления
в редакцию: 05.12.2020