ИНФОРМАТИКА И ТЕХНИЧЕСКИЕ НАУКИ
COMPUTER SCIENCE AND ENGINEERING SCIENCES

МОДЕЛИРОВАНИЕ ПРОЦЕССОВ И УПРАВЛЕНИЕ В ТЕХНИЧЕСКИХ СИСТЕМАХ
MODELLING OF PROCESSES AND MANAGEMENT IN TECHNICAL SYSTEMS

УДК 681.518

А. А. Королёв, Д. С. Карпович, Т. П. Фокин
Белорусский государственный технологический университет

КЛАССИФИКАЦИЯ ТИПОВ ДАННЫХ, ИСПОЛЬЗУЕМЫХ В АНАЛИТИКЕ БОЛЬШИХ ДАННЫХ

Статья посвящена основным категориям данных, применяемых в современных системах Big Data. В статье проводится классификация данных на структурированные, полуструктурированные и неструктурированные с описанием особенностей каждого типа, анализу их ключевых характеристик, а также сравнению их преимуществ и недостатков. Также в статье приведено описание больших данных, рассмотрены общие понятия и особенности больших данных, сферы их применения, источники их появления. Проведен анализ основных типов данных, используемых в системах Big Data, посредством сравнения их между собой по таким характеристикам, как гибкость, масштабируемость, сложность анализа, сложность интеграции, вид представления данных, эффективность хранения, сложность запросов для выборки данных. Рассматриваются технологии, которые обеспечивают эффективное использование различных типов данных в рамках Big Data, а также роль этих технологий в улучшении процессов обработки и принятия решений на предприятиях. Подчеркнута важность анализа каждого из существующих типов данных для оптимизации бизнес-операций.

Ключевые слова: большие данные, структурированные данные, полуструктурированные данные, неструктурированные данные.

Для цитирования: Королёв А. А., Карпович Д. С., Фокин Т. П. Классификация типов данных, используемых в аналитике больших данных // Труды БГТУ. Сер. 3, Физико-математические науки и информатика. 2025. № 1 (290). С. 31–35.
DOI: 10.52065/2520-6141-2025-290-6.

A. A. Korolyov, D. S. Karpovich, T. P. Fokin
Belarusian State Technological University

CLASSIFICATION OF DATA TYPES USED IN BIG DATA ANALYTICS

The article is devoted to the main categories of data used in modern Big Data systems. The article classifies data into structured, semi-structured and unstructured with a description of the features of each type, an analysis of their key characteristics, and a comparison of their advantages and disadvantages. The article also describes big data, considers general concepts and features of big data, areas of their application, and sources

of their appearance. An analysis of the main types of data used in Big Data systems is carried out by comparing them with each other according to such characteristics as flexibility, scalability, complexity of analysis, complexity of integration, type of data representation, storage efficiency, complexity of queries for data selection. The technologies that ensure the efficient use of various types of data within Big Data are considered, as well as the role of these technologies in improving the processes of processing and decision-making at enterprises. The importance of analyzing each of the existing data types for optimizing business operations is emphasized.

Keywords: big data, structured data, semi-structured data, unstructured data.

For citation: Korolyov A. A., Karpovich D. S., Fokin T. P. Classification of data types used in big data analytics. *Proceedings of BSTU, issue 3, Physics and Mathematics. Informatics*, 2025, no. 1 (290), pp. 31–35 (In Russian).

DOI: 10.52065/2520-6141-2025-290-6.

Введение. В промышленной и информационной индустрии данные играют центральную роль. С развитием более быстрых сетей, более широкого пространства для хранения и новых сенсорных технологий компании получают доступ к все большим объемам информации, и иногда этот объём оказывается даже слишком большим.

Термин Big Data описывает огромные объемы данных, как структурированных, так и неструктурированных, собираемых каждый день предприятиями. Большие данные могут поступать из любого количества источников – от социальных сетей и поиска Google до данных, собранных датчиками с промышленного оборудования.

Большие данные отличаются от других наборов данных тем, что из-за их размера их сложно обрабатывать с использованием традиционных методов обработки данных. Большие данные определяются с применением следующих четырех измерений – 4V больших данных [1, 2]:

– объем: объем (структурированных и неструктурированных) данных, сгенерированных и сохраненных;

– разнообразие: различные типы доступных данных, такие как текст, изображения, видео и т. д.;

– достоверность: достоверность и качество данных, включая точность, согласованность и полноту;

– скорость: скорость, с которой компания генерирует и обрабатывает данные.

Сферы применения больших данных

Чтобы получить полное представление о больших данных, необходимо изучить различные особенности и типы больших данных, а также то, какой вклад они вносят в науку о данных.

Большие данные можно условно разделить на три основных типа:

– структурированные данные;

– полуструктурированные данные;

– неструктурированные данные.

В зависимости от типа структуры данных применяются различные подходы для получения информации, требуемой из различных структур данных.

Основная часть. *Структурированные данные.* Структурированные данные в Big Data можно определить как отформатированные, четко определенные данные, которые следуют общепринятым правилам. В отличие от неструктурированных данных, они поступают в виде схемы, которая может быть представлена в табличной форме.

Различные графики, такие как столбчатые диаграммы, круговые диаграммы и т. д., создаются только из структурированных данных. Они также известны как количественные данные, поскольку значения в структурированных данных представлены в количественном выражении.

Для управления структурированными данными в среде больших данных используется SQL [3].

Согласно определению структурированных данных, это не что иное, как организованные данные, которые хранятся в базах данных, наборах данных и электронных таблицах. Он был изобретен ученым IBM Эдгаром Коддом и используется такими компаниями, как IBM, Microsoft, Oracle и др. Структурированные данные играют важную роль в развитии среды больших данных, поэтому они используются во всем мире на ежедневной основе. Они имеют следующие характеристики:

– простота использования: самое большое преимущество структурированных данных в больших данных заключается в том, что они делают данные пригодными для использования

даже среднестатистическим бизнес-пользователем, так как нет необходимости иметь подробную информацию о различных типах данных и их взаимосвязях;

– структурное представление: данные хранятся как в столбцах, так и в строках, и это легко позволяет обеспечить безопасность данных, и они также имеют определенную структуру, которая помогает в легком хранении и доступе к данным;

– каждая из таблиц имеет определенный атрибут: таблицу можно настраивать, что включает обновление, чтение и удаление или добавление новых данных, а это позволяет производить беспроблемные операции, зачастую выполняемые в реляционной модели при помощи языка структурированных запросов (SQL).

Неструктурированные данные. Неструктурированные данные относятся к нетрадиционной модели, в которой нельзя применять заранее определенные правила. Аналогично, в среде данных объем неструктурированных данных составляет до 90% данных, собранных с разных предприятий [4].

Тексты, аудио, видео, изображения – все это неструктурированные данные. В отличие от структурированных, в больших данных они не имеют фиксированных параметров, поэтому еще одно название неструктурированных данных – качественные данные.

Из изображений или видео возможно взять информацию, чтобы узнать ее качество и получить обратную связь с ней. Они отражены в сравнении с другими типами данных в таблице.

Сравнение типов данных, используемых в Big Data

Аспект	Данные		
	структурированные	неструктурированные	полуструктурированные
Описание	Организованные	Отсутствие предопределенной структуры	Сочетание организованности и гибкости
Примеры	Операции продаж в реляционной базе данных, записи студентов в электронных таблицах	Посты в социальных сетях, отзывы клиентов, медицинские изображения и аудиозаписи	XML-документы, JSON данные, базы данных NoSQL
Эффективность хранения	Эффективное хранение и поиск благодаря организованному формату	Различная эффективность хранения в зависимости от типов контента, возможны сложности в управлении ими	Сочетают в себе эффективность хранения и гибкость, оптимизированы для сложных структур данных
Запросы	Хорошо подходят для структурированных языков запросов (SQL), эффективное выполнение запросов	Сложность запросов, требуют применения передовых методов, таких как обработка естественного языка	Требуются специализированные методы запросов, адаптируемые к сложным отношениям
Сложность данных	Хорошо организованы и просты в управлении	Хаотичны и сложны в организации из-за отсутствия структуры	Сочетают гибкость с определенным уровнем организованности и умеренной сложностью
Гибкость	Ограниченная гибкость, данные должны соответствовать предопределенной структуре	Очень гибкие, могут охватывать разнообразный контент, но могут быть лишены единообразия	Обеспечивают гибкость, сохраняя при этом определенный уровень структуры, адаптируемой к изменениям
Интеграция	Хорошо подходят для традиционных реляционных баз данных и структурированных приложений	Могут потребоваться передовые методы интеграции из-за разнообразия форматов	Адаптируемы для веб-приложений, API и систем с различными источниками данных
Сложность анализа	Проще анализировать, подходят для количественного анализа и отчетности	Требуются передовые методы анализа настроений, распознавания образов и т. д.	Сложный анализ может затрагивать специализированные методы, но учитываются разнообразные структуры данных
Масштабируемость	Эффективны для управления большими объемами данных благодаря структурированному формату	Проблемы масштабируемости из-за разнообразия данных	Масштабируемы, но сложность может увеличиваться с объемом и структурой данных

Для обработки неструктурированных данных используются NoSQL базы данных, поскольку для них нет жесткой и быстрой модели. В использовании неструктурированных данных есть как важные преимущества, так и недостатки, которые необходимо учитывать заранее при возможной обработке этих данных. Одним из преимуществ использования неструктурированных данных является большее количество информации и качественных аспектов, которые структурированные данные могут упускать из виду. Также разнообразная природа неструктурированных данных более точно отражает реальные сценарии и может быть ценной для принятия решений и анализа тенденций. Наряду с этим неструктурированные данные подпитывают инновации в областях распознавания изображений и машинного обучения.

Полуструктурированные данные. Полуструктурированные данные относятся к типу данных, которые находятся где-то между структурированными данными, используемыми традиционными реляционными базами данных, и неструктурированными данными, такими как файлы мультимедиа и изображения. Хотя они не вписываются в таблицы с предопределенной схемой, как структурированные данные, полуструктурированные данные поддерживают некоторую организацию или иерархию, например, метаданные или семантические теги, что делает их доступными для поиска с помощью запросов, в отличие от неструктурированных данных. Например, XML- и JSON-документы, базы данных NoSQL являются основными вариантами представления и хранения таких данных [5]. Добавление структуры делает полуструктурированные данные более полезными, чем чисто неструктурированные данные, для предприятий, которые все больше полагаются на данные для принятия решений.

Полуструктурированные данные могут поступать из самых разных источников: от устройств и датчиков Интернета вещей (IoT) до веб-страниц, электронных писем и многого другого. Организациям необходимо иметь возможность извлекать информацию из всех своих данных, независимо от типа.

Так как полуструктурированные данные – это гибрид структурированных и неструктурированных данных, по этой причине они разделяют некоторые аспекты с обоими типами. Они не так жестко структурированы, как первые, но содержат идентификационную информацию или теги, которые делают их более доступными для поиска и действий, чем вторые. Организации собирают полуструктурированные данные и создают их, добавляя информацию к неструктурированным данным.

Важность полуструктурированных данных заключается в том, что неструктурированные данные составляют более 90% всех данных, генерируемых в мире, и ежегодно их количество растет на 55%, и, поскольку предприятиям, собирающим огромные объемы неструктурированных данных, необходимо сделать их пригодными для использования, они могут добавлять теги или информацию, чтобы превратить неструктурированные данные в полуструктурированные для удовлетворения этой потребности. В противном случае они упустят потенциальные идеи из большей доли данных, которые они собирают и хранят.

Заключение. В проведенной работе подчеркивается важность правильной классификации данных для эффективного анализа и принятия решений. Понимание различий между структурированными, полуструктурированными и неструктурированными данными позволяет выбрать наиболее подходящие инструменты и методы обработки. В статье акцентируется внимание на преимуществах и недостатках каждого из видов данных, проведено сравнение трех видов данных на основе таких характеристик, как гибкость, масштабируемость, сложность анализа, сложность интеграции, вид представления данных, эффективность хранения данных, сложность запросов для выборки данных. Из проведенного анализа можно сделать вывод, что в условиях роста объема и сложности данных важно продолжать совершенствовать методы их обработки и интеграции для максимальной эффективности в аналитике больших данных.

Список литературы

1. Chen M., Mao S., Liu Y. Big data: A survey // Mobile Networks and Applications. 2014. Vol. 19 (2). P. 171–209. DOI: 10.1007/s11036-013-0489-0.

2. Katal A., Wazid M., Goudar R. H. Big Data: Issues, challenges, tools and good practices // 6th International Conference on Contemporary Computing (IC3). 2013. P. 404–409. DOI: 10.1109/IC3.2013.6612229.

3. Chen C. P., Zhang C. Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data // Information Sciences. 2014. Vol. 275. P. 314–347. DOI: 10.1016/j.ins.2014.01.015.

4. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics // International Journal of Information Management. 2015. Vol. 35 (2). P. 137–144. DOI: 10.1016/j.ijinfomgt.2014.10.007.

5. Papakonstantinou Y. Semistructured Models, Queries and Algebras in the Big Data Era // SIGMOD '16: Proceedings of the 2016 International Conference on Management of Data. 2016. P. 2229–2233. DOI: 10.1145/2882903.2912573.

References

1. Chen M., Mao S., Liu Y. Big data: A survey. *Mobile Networks and Applications*, 2014, vol. 19 (2), pp. 171–209. DOI: 10.1007/s11036-013-0489-0.

2. Katal A., Wazid M., Goudar R. H. Big Data: Issues, challenges, tools and good practices. *6th International Conference on Contemporary Computing (IC3)*, 2013, pp. 404–409. DOI: 10.1109/IC3.2013.6612229.

3. Chen C. P., Zhang C. Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, 2014, vol. 275, pp. 314–347. DOI: 10.1016/j.ins.2014.01.015.

4. Gandomi A., Haider M. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 2015, vol. 35 (2), pp. 137–144. DOI: 10.1016/j.ijinfomgt.2014.10.007.

5. Papakonstantinou Y. Semistructured Models, Queries and Algebras in the Big Data Era. *SIGMOD '16: Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 2229–2233. DOI: 10.1145/2882903.2912573.

Информация об авторах

Королёв Артём Андреевич – старший преподаватель кафедры автоматизации производственных процессов и электротехники. Белорусский государственный технологический университет (ул. Свердлова, 13а, 220006, г. Минск, Республика Беларусь). E-mail: korolev@belstu.by

Карпович Дмитрий Семёнович – кандидат технических наук, доцент, заведующий кафедрой автоматизации производственных процессов и электротехники. Белорусский государственный технологический университет (ул. Свердлова, 13а, 220006, г. Минск, Республика Беларусь). E-mail: d.karpovich@belstu.by

Фокин Тимофей Павлович – преподаватель-стажер кафедры автоматизации производственных процессов и электротехники. Белорусский государственный технологический университет (ул. Свердлова, 13а, 220006, г. Минск, Республика Беларусь). E-mail: fokin@belstu.by

Information about the authors

Korolyov Artyom Andreevich – Senior Lecturer, the Department of Automation of Production Processes and Electrical Engineering. Belarusian State Technological University (13a Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: korolev@belstu.by

Karpovich Dzmitry Semenovich – PhD (Engineering), Associate Professor, Head of the Department of Automation of Production Processes and Electrical Engineering. Belarusian State Technological University (13a Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: d.karpovich@belstu.by

Fokin Timophej Pavlovich – teacher trainee, the Department of Automation of Production Processes and Electrical Engineering. Belarusian State Technological University (13a Sverdlova str., 220006, Minsk, Republic of Belarus). E-mail: fokin@belstu.by

Поступила после доработки 15.01.2025