УДК 004.6; 003.338 DOI: 10.18324/2224-1833-2022-1-43-48

«Большие» данные и их роль в решении бизнес-задач

М.Ю. Иванов[1a], В.В. Надршин[2b], М.А. Полячкова[1c]

[1]Братский государственный университет, ул. Макаренко, 40, Братск, Россия
[2]Иркутский национальный исследовательский технический университет, ул. Лермонтова, 83, Иркутск, Россия
[a]nis@brstu.ru, [b]nadrshin@istu.edu, [c]mpolyachkova@mail.ru
Статья поступила 10.01.2022, принята 16.02.2022

Представлены результаты исследований технологий «больших» данных: проанализирована и уточнена терминология с учётом специфики российского IT-рынка, раскрыты признаки «больших» данных (объем, скорость обновления, разнообразие, изменчивость, значение данных). Выявлены основные источники и принципы работы с «большими» данными (расширяемость системы, устойчивость к отказу, локализация). Предложены основные способы анализа «больших» данных (глубинный анализ и классификация данных, краудсорсинг, сплит-тестирование, прогнозирование, машинное обучение, анализ сетевой активности). Сформулирована разница в подходах к анализу данных с помощью традиционных инструментов и механизмов «больших» данных. Показана возможность практического использования технологий «больших» данных в экономике: определены субъекты (поставщики инфраструктуры, дата-майнеры, системные интеграторы, заказчики, разработчики. Определены перспективы развития (облачные хранилища, «тёмные» данные, блокчейн, системы самообслуживания), драйверы и ограничители «больших» данных, приведены основные преимущества применения «больших» данных для повышения эффективности работы коммерческих предприятий в сфере продажи товаров и услуг.

Ключевые слова: массивы данных, глубинный анализ и классификация данных, краудсорсинг, машинное обучение, инсайт, дата-майнинг, облачные хранилища, блокчейн, социальные сети.

«Big» data and its role in solving business problems

M.Yu. Ivanov[1a], V.V. Nadrshin[2b], M.A. Polyachkova[1c]

[1]Bratsk State University; 40, Makarenko St., Bratsk, Russia
[2]Irkutsk National Research Technical University; 83, Lermontov St., Irkutsk, Russia
[a]nis@brstu.ru, [b]nadrshin@istu.edu, [c]mpolyachkova@mail.ru
Received 10.01.2022, accepted 16.02.2022

The results of research on technologies of «big» data are presented: the terminology, taking into account the specifics of the Russian IT market, is analyzed and refined. The signs of «big» data (volume, velocity, variety, variability, value) are revealed. The main sources and principles of working with «big» data (system extensibility, resistance to failure, localization) are identified. The main methods of analyzing «big» data (deep analysis and data classification, crowdsourcing, split testing, forecasting, machine learning, analysis of network activity) are proposed. The difference in approaches to data analysis using traditional tools and mechanisms of «big» data is formulated. The possibility of practical use of «big» data technologies in the economy is shown: the subjects (infrastructure providers, data miners, system integrators, customers, developers) are identified. Development prospects (cloud storage, «dark» data, blockchain, self-service systems), drivers and limiters of «big» data are determined. The main advantages of using «big» data to improve the efficiency of commercial enterprises in the sale of goods and services are shown.

Keywords: data sets, deep analysis and data classification, crowdsourcing, machine learning, insight, data mining, cloud storage, blockchain, social networks.

Введение. Сама по себе концепция «больших» данных (далее – БоД) не нова, поскольку большие наборы данных начали использовать еще в 60-70-ые годы прошлого века с появлением центров обработки данных и реляционных баз данных.

В 2001 году учёные консалтинговой компании «Gartner» (г. Стэмфорд, США), специализирующейся на рынках информационных технологий, назвали термином БоД разнообразную информацию, которая поступает с постоянно растущей скоростью и объем которой также постоянно увеличивается.

Иными словами, БоД – это необычайно большие и сложные наборы данных разных форматов

представления, объемы которых при этом постоянно увеличиваются. Зачастую, традиционное программное обеспечение не может справиться с такими данными, так как их размер настолько велик, что [1-5]. Вместе с тем, БоД уже используются во многих отраслях от маркетинга до урбанистики. Широко используют их и для бизнес-задач, решение которых раньше являлось слишком сложным или экономически нецелесообразным.

В 2008 году понятие БоД использовал редактор журнала «Nature» в специальном выпуске, посвященном взрывному росту мировых потоков информации, освоить которые могут помочь новые инструменты и более развитые технологии. Так, Клиффорд Линч предложил считать под БоД любые массивы неоднородных данных объемом свыше 150 Гб в сутки [6].

Разумеется, одним из крупнейших производителей БоД была и остается наука. Так, один только Большой адронный коллайдер производит более 15 ПБ (10^{15} байт) данных в год, а NASA хранит на своих серверах более 37 ПБ данных об изменении климата. Учёные всего мира в последние годы работают над механизмами, значительно повышающими скорость операций с данными. Справиться с этими задачами может помочь учёт семантики данных не только во время обработки, но и при организации их хранения. Новый подход к хранению позволяет сократить время на операции с данными более чем в два раза, а также отказаться от использования дополнительных вычислительных мощностей.

Согласно статистике ведущих аналитических агентств, в 2005 году мир оперировал 4-5 ЭБ (10^{18} байт) информации. Через 5 лет объемы БоД выросли до 0,19 ЗБ (10^{21} байт). В 2012 году эти показатели возросли до 1,8 ЗБ, а в 2015 – до 7 ЗБ. После 2020 года системы БоД оперируют уже 42-45 ЗБ информации [7].

До 2011 года БоД практического применения не имели и рассматривались только как направление научной деятельности. Однако уже в начале 2012 года проблема обработки огромных массивов неструктурированной и неоднородной информации стала актуальной из-за растущих с большой скоростью объемов данных. Именно в этот период к развитию БоД подключились такие мастодонты цифрового бизнеса, как «Microsoft», «IBM», «Oracle» и др. С 2014 года принципы организации БоД стали вводить в образовательные программы высшего образования и внедрять в прикладные науки: инженерию, физику, социологию.

Теоретические основы БоД. БоД обычно определяют довольно просто – это огромный объем информации, часто бессистемной, которая хранится на каком - либо цифровом носителе [7]. Однако массивы данных с приставкой «big» настолько велики, что привычными средствами аналитики и структурирования обработать их невозможно. Поэтому под термином БоД следует понимать не только сами данные, как это принято за рубежом, но и технологии поиска, применения и обработки неструктурированной информации в больших объёмах.

БоД достаточно эффективно обрабатываются с помощью масштабируемых программных продуктов, которые появились в конце 2000-х годов и стали альтернативой традиционным системам управления базами данных и решениям «business intelligence» [8-9].

Первоначально БоД характеризовались следующими признаками:

- объем информации. Данные измеряются по занимаемому пространству на цифровом носителе. При работе с большими данными требуется обрабатывать внушительные объемы неструктурированных данных низкой плотности. Как отмечалось выше, к БоД относят массивы данных свыше 150 Гб в сутки;

- скорость обновления информации. Информация регулярно обновляется и для обработки в режиме реального времени уже необходимы интеллектуальные технологии БоД;

- разнообразие и разнородность. Доступные данные принадлежат к разным типам. Традиционные типы данных структурированы и могут быть сразу сохранены в реляционной базе данных. С появлением больших данных данные стали поступать в неструктурированном виде. Такие неструктурированные и полуструктурированные типы данных как текст, аудио и видео требуют дополнительной обработки для определения их значения и поддержки метаданных[7].

В настоящее время специфика БоД требует введения двух дополнительных признаков:

- изменчивость. Потоки данных могут иметь спады и пики, периодичность, сезонность. Такие всплески неструктурированной информации требуют мощных инструментов для обработки и сложны в управлении;

- значение данных. Задачу управления БоД усложняет то, что информация может иметь разную сложность для восприятия и переработки. В связи с этим интеллектуальные системы управления БоД должны уметь определять степень важности поступающей информации, для того чтобы быстро ее структурировать.

Технологии работы с БоД основаны на макси-

мальном информировании пользователя о какомлибо предмете или явлении, помощи принятия верного решения взвесив все «за» и «против». В интеллектуальных системах управления БоД на основе массива информации строится модель развития, имитируются различные варианты ее поведения и отслеживаются результаты. На практике при тестировании идеи, предположения или решении проблемы запускаются миллионы подобных симуляций.

К источникам БоД относят ресурсы информационно-телекоммуникационной сети «Интернет» (блоги, социальные сети, сайты, СМИ, форумы, а также поисковые системы, обладающие достаточной технологической базой для создания новых сервисов; корпоративную информацию (архивы, транзакции, базы данных); показания всевозможных считывающих устройств (метеорологических приборов, оборудования сотовой и спутниковой связи, датчиков, маяков, камер и т.д.) [10].

Принципы работы с БоД включают три основных фактора:

- расширяемость системы, под которой обычно понимают горизонтальную масштабируемость носителей информации. То есть, при возрастании объемов поступающих данных увеличивается и мощность, и количество серверов для их хранения;

- устойчивость к отказу. Повышать количество цифровых носителей и интеллектуальных систем соразмерно объемам поступающих данных можно до бесконечности. Но это не означает, что часть устройств не будет выходить из строя и устаревать. Поэтому одним из факторов стабильной работы с БоД является отказоустойчивость серверов;

- локализация. Отдельные массивы БоД целесообразно хранить и обрабатывать в пределах одного выделенного сервера, что позволяет экономить время, ресурсы и оптимизировать расходы на передачу данных [11].

К основным способам анализа БоД относят:

- глубинный анализ и классификация данных. Эти методы применялись при работе и с обычной структурированной информацией в небольших объемах. Однако в новых условиях используются усовершенствованные математические алгоритмы, основанные на достижениях в цифровой сфере;

- краудсорсинг. В основе этой технологии заложена возможность получения и обработки потоков в миллиарды байтов из множества источников. Конечное число источников БоД в данном случае может быть ограничено лишь мощностью системы;

- сплит-тестирование. Из массива данных выбираются несколько элементов, сравнивая их между собой поочередно до и после изменения. Тесты позволяют определить те факторы, которые оказывают наибольшее влияние на выбранные элементы. Например, с помощью сплит-тестирования можно провести огромное количество итераций, постепенно приближаясь к достоверному результату;

- прогнозирование. Системе стараются заранее задать те или иные параметры и в дальнейшем проверяют поведение выбранного объекта или процесса на основе поступления больших массивов информации;

- машинное обучение. Системы искусственного интеллекта в перспективе способны поглощать и обрабатывать большие объемы несистематизированных данных, используя их впоследствии и для самостоятельного обучения;

- анализ сетевой активности. Технологии БоД могут использоваться для исследования структуры социальных сетей, взаимоотношений между владельцами аккаунтов, группами, сообществами. На основе этой информации создаются целевые аудитории по интересам, геолокации, возрасту и прочим метрикам [12].

В таблице 1 показана разница в подходах к анализу данных с помощью традиционных инструментов и технологий БоД.

Таблица 1. Разница подходов к анализу данных

Подход / № п/п	Традиционная аналитика	БоД-аналитика
1	Постепенный анализ небольших пакетов данных	Обработка сразу всего массива доступных данных
2	Сортировка и редактирование данных перед обработкой	Данные обрабатываются в их исходном виде
3	Старт с гипотезы и ее тестирования относительно данных	Поиск корреляций по всем данным до получения искомой информации
4	Данные подлежат сбору, обработке, хранения и лишь затем анализируются	Анализ и обработка данных в режиме реального времени по мере поступления

С учётом подходов к анализу данных, представленных в таблице 1, можно сделать вывод о том, что технологии БоД позволяют получить из большого объема исходных данных новую, ранее неизвестную информацию. Подобное решение часто называют инсайтом, что означает открытие, озарение, догадку, внезапное понимание.

Практическое использование технологий БоД в экономике. Известно, что стратегии развития бизнеса, маркетинговые мероприятия, реклама основаны на анализе и работе с имеющимися данными [13]. Логично предположить, что технологии БоД позволяют переработать огромные объемы данных и, соответственно, максимально точно скорректировать направление развития бренда, продукта или услуги.

Все субъекты, имеющие дело с БоД, можно условно разделить на следующие группы:

- поставщики инфраструктуры, решающие задачи предобработки и хранения данных;

- дата-майнеры, разрабатывающие алгоритмы, помогающие заказчикам извлекать ценные сведения из непрерывных потоков данных;

- системные интеграторы, которые внедряют системы анализа БоД на стороне заказчика;

- заказчики из отраслей финансов, телекоммуникаций, ритейла, приобретающие программно-аппаратные комплексы и алгоритмы работы с БоД;

- разработчики, которые предлагают готовые решения (сервисы) на основе доступа к БоД [14].

Так, системы анализа БоД собирают информацию об интересах пользователей и анализируют её. Сервисы БоД на основе поведения отдельно взятого пользователя могут перестроить личный контент какого-либо сайта: настроить размещение услуг и товаров в каталогах, создать персонализированные и таргетированные почтовые рассылки, то есть, эффективно предлагать коммерческие предложения выделенной целевой аудитории, а не всем подряд [15].

Сервисы БоД по управлению закупками цифровой рекламы помогают эффективно участвовать в RTB-аукционах, используя кросс-канальный, поисковый и товарный ретаргетинг для привлечения нужных покупателей.

Перспективами развития на ближайшие годы можно считать интеграцию БоД в средний и малый бизнес и крайне популярные сегодня стартапы. Технологии БоД в указанных выше сферах реализуются следующим образом:

- облачные хранилища. Технологии хранения и работы с данными в он-лайн-пространстве гораздо дешевле содержания дата-центра [16];

- «тёмные» данные, предусматривающие сбор и хранение не оцифрованных данных предприятия, которые не имеют значимой роли для развития бизнеса, но нужны, к примеру, на техническом и законодательном уровнях;

- блокчейн – распределенная база данных, содержащая информацию в виде цепочки блоков, в каждом из которых записано определенное число всех транзакций, проведенных участниками системы. Соответственно, упрощение интернет-транзакций позволяет снизить затраты на проведение этих операций;

- системы самообслуживания, представляющие собой специальные платформы для среднего и малого бизнеса. Такие системы внедряются с 2016 года и позволяют предприятию самостоятельно хранить и систематизировать данные [17].

Некоторые компании активно применяют БоД для прогнозирования потребительского спроса. Технологии БоД классифицируют ключевые атрибуты существующих и снятых с производства продуктов и услуг и моделируют связи между этими атрибутами и коммерческим успехом предложений. На основе полученных данных создаются предиктивные модели для новых продуктов и услуг. Кроме того, указанная фирма использует данные и статистику, получаемые от фокусных групп, а также из социальных сетей по результатам рыночных тестов и пробных продаж [18].

В целом, технологии БоД предоставляют предприятиям следующие преимущества:

- разработка проектов, которые с высокой вероятностью станут востребованными у покупателей;

- упрощается планирование и увеличивается скорость запуска новых проектов;

- ускоряется взаимодействие с клиентами и контрагентами;

- снижаются издержки в работе с поставщиками и клиентами;

- изучаются и анализируются требования клиентов с существующим сервисом предприятия, корректируется работа обслуживающего персонала;

- выявляются лояльные и неудовлетворенные клиенты за счет анализа разнообразной информации из блогов, социальных сетей и других источников БоД;

- благодаря аналитической работе с БоД привлекается и удерживается целевая аудитория предприятия.

Следует отметить, что интерес к технологиям БоД в России, несомненно, растет, но на российском рынке у БоД есть как драйверы, так и ограничители (таблица 2).

Таблица 2. Драйверы и ограничители технологий БоД

Драйверы	Ограничители
Высокий спрос на БоД для повышения конкурентоспособности	Необходимость обеспечивать безопасность и конфиденциальность данных
Развитие методов обработки медиафайлов на мировом уровне	Нехватка квалифицированных кадров
Реализация отраслевых планов по импортозамещению программного обеспечения	В большинстве российских компаний объем накопленных информационных ресурсов не достигает уровня БоД
Тренд на использование услуг российских провайдеров и системных интеграторов	Новые технологии сложно внедрять в традиционные информационные системы компаний
Создание технопарков, способствующих развитию информационных технологий	Высокая стоимость информационных технологий
Государственные программы по внедрению грид-систем – виртуальных суперкомпьютеров, которые распространяются по кластерам и связываются в сеть	Заморозка инвестиционных проектов в России и отток зарубежного капитала
Перенос на территорию России серверов, обрабатывающих персональную информацию	Рост цен на импортную продукцию и услуги

Заключение. БоД имеют долгую историю развития, однако их потенциал еще далеко не раскрыт. Так, облачные технологии обеспечивают по-настоящему гибкие возможности масштабирования, что позволяет разработчикам развертывать кластеры для тестирования выборочных данных по требованию и раздвигать границы применения БоД еще шире [19].

С появлением Интернета вещей все большее число устройств получает доступ к сети, что позволяет собирать данные о моделях действий пользователей и работе различных продуктов [20]. А с развитием технологий машинного обучения объем данных вырос еще больше.

При этом нельзя утверждать, что есть отдельные виды БоД. Суть БоД состоит в том, что объединяются самые различные типы данных, из которых извлекается новая, ранее недоступная информация. Предоставляя большее количество информации, БоД дают возможность получать более полные ответы на поставленные вопросы. Более подробные ответы увеличивают уверенность в достоверности данных, что, в свою очередь, обеспечивает абсолютно новые подходы к решению бизнес-задач.

Литература

1. Иванов М.Ю., Косякова В.В. Критерии качества программного обеспечения // Труды Братского гос. ун-та. Сер. Проблемы управления социально-экон. развитием регионов Сибири. Братск: Изд-во БрГУ, 2011. С. 71-74.
2. Иванов М.Ю., Косякова В.В. Современные аспекты управления качеством программного обеспечения // Труды Братского гос. ун-та. Сер. Проблемы управления социально-экон. развитием регионов Сибири. Братск: Изд-во БрГУ, 2011. С. 75-78.
3. Иванов М.Ю. Современные аспекты разработки программного обеспечения экономико-управленческих систем и процессов // Системы. Методы. Технологии. 2013. № 1 (17). С. 145-148.
4. Иванов М.Ю. Автоматизация сетевого планирования и управления // Системы. Методы. Технологии. 2013. № 2 (18). С. 63-69.
5. Иванов М.Ю. Современные аспекты эффективности программного обеспечения // Труды Братского гос. ун-та. Сер. Проблемы управления социально-экон. развитием регионов Сибири. Братск: Изд-во БрГУ, 2013. С. 261-264.
6. Lynch Clifford A. Big data: How do your data grow? // Nature. 2008. V. 455. P. 7209.
7. Smolan R., Erwitt J. The human face of big data. Sausalito: Against All Odds Productions, 2012. 223 p.
8. Alchinov A.I., Tavbulatova Z.K., Dudareva O.V., Ivanov M.Yu. Modern approach to enterprise information systems // Journal of Physics: Conference Series. 2020. V. 1661. P. 012164.
9. Malsagov B.S., Ivanov M.Yu., Natalevich L.F. Structural features of accounting automation application // Journal of Physics: Conference Series: International Conference on IT in Business and Industry (ITBI 2021). 2021. V. 2032. P. 012128.
10. Daudov I.M., Sygotina M.V., Nadrshin V.V. Principles of the transition from 4G LTE to 5G // Journal of Physics: Conference Series: International Conference on IT in Business and Industry (ITBI 2021). 2021. V. 2032. P. 012006.
11. Zhigalov K., Kuznetsova S.Y., Sygotina M.V. Development of functional fault-tolerant system. 2020. V. 1661. P. 012166.
12. Daudov I.M., Gavrilova Zh.L., Kudashkin V.A. Liable Bluetooth tracking technology for enhancement of location-based services // IOP Conference Series: Materials Science and Engineering. 2021. V. 1111. P. 012043.
13. Alchinov A.I., Yalovenko O.V., Sygotina M.V. Main architectural patterns of web applications and web services using the example of banking systems // Journal of Physics: Conference Series: International Conference on IT in Business and Industry (ITBI 2021). 2021. V. 2032. P. 012116.
14. Han J., Kamber M., Pei J. Data mining: concepts and techniques. Waltham: Morgan Kaufmann Publishers, 2011. 703 p.

15. Turluyev R.R., Ivanov M.Yu., Beregova G.M. Models and classifications of secure resource management methods in distributed information communication networks // IOP Conference Series: Materials Science and Engineering. 2021. V. 1111. P. 012062.
16. Tavbulatova Z.K., Zhigalov K., Kuznetsova S.Y., Patrusova A.M. Types of cloud deployment. 2020. V. 1582. P. 012085.
17. Tan Pang-Ning, Steinbach M., Kumar V. Introduction to Data Mining. London: Pearson Education Limited, 2014. 732 p.
18. Vakhrusheva M.Yu., Khaliev M.S.-U., Pokhomchikova E.O. Barclays' application of information system in manufacturing process // Journal of Physics: Conference Series: International Conference on IT in Business and Industry (ITBI 2021). 2021. V. 2032. P. 012129.
19. Dudareva O.V., Alchinov A.I., Vakhrusheva M.Yu. Basic principles for data protection for decision-making and control systems // Journal of Physics: Conference Series: International Conference on IT in Business and Industry (ITBI 2021). 2021. V. 2032. P. 012078.
20. Daudov Kh.A., Patrusova A.M., Nadrshin V.V. Analysis of narrowband data transfer technologies on the Internet of Things (IoT) // IOP Conference Series: Materials Science and Engineering. 2021. V. 1111. P. 012016.