УДК 519.683: 330.4

Методы анализа больших данных

В статье дано определение термину «большие данные», выявлены особенности больших данных, описаны методы их анализа, раскрыты отличия традиционных способов обработки данных от технологий больших данных, определена значимость применения больших данных предприятиями.

О.В. ДОМАКУР,
кандидат экономических наук,
доцент, ученый секретарь

Белорусская государственная академия связи

Ключевые слова:
большие данные, методы анализа больших данных, технологии обработки больших данных.

Введение. На рубеже десятилетий эпохи цифровой трансформации экономики, ознаменовавшей этап зрелости информационного общества, технологии больших данных кардинально модифицируют производственные и управленческие процессы, заменяют многочасовой интеллектуальный труд человека и развивают искусственный интеллект, способствуют ускорению обработки постоянно возрастающих огромных объемов данных о жизнедеятельности человека и хозяйственной деятельности предприятий, находят ранее скрытые зависимости, упрощают восприятие результатов обработки данных человеком, дают рекомендации для принятий решений. В условиях тотального изобилия постиндустриального общества именно технологии больших данных позволяют индивидуализировать производство, предугадать и удовлетворить желания потребителей, при этом обеспечить приемлемые цены, произвести общественные блага без участия государственного финансирования. Научное изучение этой темы позволит в дальнейшем выработать методические материалы для обучения не только технических специалистов, но и филологов, финансистов, маркетологов, управленцев.

Определение и особенности больших данных. Термин «большие данные» рассматривают с точки зрения объемов данных и методов их обработки для целей эффективного использования. Он обозначает наборы данных, размер которых превосходит возможности типичных баз данных (БД) по занесению, хранению, управлению и анализу информации, с учетом роста объемов этих данных [3].

Исследователи отмечают бесполезность самих данных до тех пор, пока они не превратятся в необходимую для конкретных целей информацию. Поэтому применение термина «большие данные» предполагает и порой подразумевает средства анализа этих данных для получения полезной информации.

Предприятия и организации создают огромные объемы данных, однако бóльшая их часть представлена в слабоструктурированном формате (вебжурналы, видеозаписи, текстовые документы, машинный код или геопространственные данные), скорость обновления этих данных возрастает. Они хранятся во множестве разнообразных хранилищ, иногда за пределами организации. Традиционные методы анализа данных в таком виде не позволяют пользоваться ими полностью и наиболее эффективно.

Понятие больших данных подразумевает работу с данными огромного объема и разнообразного состава, часто обновляемыми и находящимися в разных источниках, в целях увеличения эффективности их использования, получения полезной информации для создания новых продуктов, повышения эффективности производственных процессов и конкурентоспособности предприятия.

Термин «большие данные» сочетает в себе как количественные и качественные характеристики данных, так и особенности их обработки для конкретных значимых целей экономической деятельности.

«Википедия» дает следующее определение: «в широком смысле о больших данных говорят как о социально-экономическом феномене, связанном с появлением технологических возможностей анализировать огромные массивы данных, в некоторых проблемных областях – весь мировой объем данных и вытекающих из этого трансформационных последствий». Аналитики компании IBS «весь мировой объем данных» оценили такими величинами:

2003 г. – 5 эксабайт данных (1 ЭБ = 1 млрд гигабайт);

2008 г. – 0,18 зеттабайта (1 ЗБ = 1024 эксабайта);

2015 г. – более 6,5 зеттабайта;

2020 г. – 40–44 зеттабайта (прогноз);

2025 г. – этот объем вырастет еще в 10 раз [1].

Консалтинговая компания Forrester дает краткую формулировку: «Большие данные объединяют техники и технологии, которые извлекают смысл из данных на экстремальном пределе практичности» [3].

«Большие данные (Big Data) – обозначение структурированных и неструктурированных данных огромных объемов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов, альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence» [1].

К 2020 году исследования выявили следующие критерии больших данных (рис. 1): объем (Volume), скорость прироста (Velocity), разнообразие (Variety), достоверность (Veracity), жизнеспособность (Viability), ценность (Value), переменчивость (Variability), визуализация (Visualization) [3].

Принципы работы с большими данными. *Горизонтальная масштабируемость* подразумевает увеличение числа вычислительных узлов, по которым распределяются эти данные, с возможностью обработки их без ухудшения производительности узлов.

Отказоустойчивость. Методы работы с большими данными должны учитывать возможность выхода из строя вычислительных узлов и предусматривать превентивные меры.

Локальность данных и их обработки. С целью экономии затрат на передачу данных больших объемов желательно проводить их обработку как можно ближе к месту их накопления [3].

Принципы работы с большими данными отличаются от традиционных методов обработки централизованных баз данных с характерной для них вертикальной моделью хранения хорошо

Таблица – Отличия традиционных баз данных и баз больших данных [3]

Характеристика	Традиционная база данных	База больших данных
Объем информации	От гигабайт до терабайт	От петабайт до эксабайт
Способ хранения	Централизованный	Децентрализованный
Структурированность данных	Высокоуровневая	Слабая или ее отсутсвие
Модель хранения и обработки данных	Вертикальная	Горизонтальная
Взаимосвязь данных	Сильная	Слабая

структурированных данных. Поэтому для работы с большими данными разрабатываются новые подходы и технологии.

Традиционный экономический анализ использовал:

- собранные данные с большим лагом, построенные факторные модели могли устаревать к моменту оценки результатов и построения прогнозов;

- факторы в модели были ограничены возможностями собрать приемлемую базу данных регулярного учета определенных показателей. В случае необходимости включения в модель нового фактора необходимо было перестраивать учет показателей на уровне бухгалтерских и статистических данных предприятий и отраслей, что вызывало большие трудности;

- упрощенные модели и усредненные показатели, так как сложные математические методы без автоматизации требовали много времени на обработку.

Применение современных способов обработки больших данных позволяет:

• анализировать данные в режиме реального времени и постоянно отслеживать динамику более точных показателей, а не средние величины;

• находить влияние новых факторов на конечный результат;

• прогнозировать поведение модели с учетом наиболее свежих данных, включая ранее не учитываемые факторы;

• визуализировать результаты моделирования и прогнозирования.

Совокупность технологий анализа больших данных осуществляют три основные группы операций обработки: 1) быстро поступающих данных в очень больших и постоянно увеличивающихся объемах, 2) как структурированных, так и слабоструктурированных данных параллельно; 3) большого разнообразия сценариев взаимосвязей поступающих данных.

Считается, что эти возможности технологий обработки больших данных позволяют выявить скрытые закономерности, ускользающие от ограниченного человеческого восприятия. Это открывает беспрецедентные возможности для оптимизации многих сфер нашей жизни: государственного управления, медицины, телекоммуникаций, финансов, транспорта, производства, а также создает базис для технологий искусственного интелекта, наиболее точно предсказывающего потребности и пожелания людей, и позволит в будущем индивидуализировать производство продукции и услуг с наименьшими издержками.

Методы анализа больших данных. Методы класса Data Mining (добыча данных, интеллектуальный анализ данных, глубинный анализ данных) – совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных знаний, необходимых для принятия решений [1]. К таким методам, в частности, относятся обучение ассоциативным правилам (association rule learning), классификация (разбиение на категории), кластерный анализ, регрессионный анализ, обнаружение и анализ отклонений и др.

Краудсорсинг – классификация и обогащение данных силами широкого неопределенного круга лиц, выполняющих эту работу без вступления в трудовые отношения.

Прогнозная аналитика (predictive analytics) – статистические методы, методы интеллектуального анализа данных, теории игр, которые на основе анализа текущего состояния и прошлого опыта выявляют шаблон поведения субъекта или группы и вероятности отклонения от него для составления предсказаний о будущих событиях, принятия решения с учетом рисков.

Имитационное моделирование (simulation) – метод, позволяющий строить модели, описывающие процессы так, как они проходили бы в действительности. Имитационное моделирование можно

рассматривать как разновидность экспериментальных испытаний.

Пространственный анализ (spatial analysis) – класс методов, использующих топологическую, геометрическую и географическую информацию, извлекаемую из данных.

Статистический анализ – анализ временных рядов, A/B-тестирование. A/B testing, split testing – метод маркетингового исследования; при его использовании контрольная группа элементов сравнивается с набором тестовых групп, в которых один или несколько показателей были изменены, чтобы выяснить, какие из изменений улучшают целевой показатель.

Смешение и интеграция данных (data fusion and integration) – набор техник, позволяющих интегрировать разнородные данные из разнообразных источников с целью проведения глубинного анализа (например, цифровая обработка сигналов, обработка естественного языка, включая тональный анализ, и др.)

Машинное обучение, включая обучение с учителем и без него, – использование моделей, построенных на базе статистического анализа или машинного обучения, для получения комплексных прогнозов на основе базовых моделей.

Искусственные нейронные сети, организуемые по принципам генетических алгоритмов, – эвристических алгоритмов поиска, используемых для решения задач оптимизации и моделирования путем случайного подбора, комбинирования и вариации искомых параметров с помощью механизмов, аналогичных естественному отбору в природе.

Распознавание образов – методы классификации и идентификации предметов, явлений, процессов, сигналов, ситуаций и тому подобных объектов, которые характеризуются конечным набором некоторых свойств и признаков.

Визуализация аналитических данных – представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и применения в качестве исходных данных для дальнейшего анализа. Очень важный этап анализа больших данных, позволяющий представить самые важные результаты в наиболее удобном для восприятия виде [1].

Заключение. Методы обработки больших данных позволяют предприятиям осуществлять цифровую трансформацию производственных и управленческих процессов, более точно предугадывать поведение потребителей, персонала, инвесторов, планировать ремонт или замену оборудования, обновление программного обеспечения, оптимизацию энергопотребления, движения транспорта, денежных и информационных потоков, повышать эффективность принимаемых решений за счет более высокой точности прогнозов влияния внешней и внутренней среды на деятельность предприятия.

ЛИТЕРАТУРА

1. Big Data. IT Enterprice. [Электронный ресурс]. – Режим доступа: https://www.it.ua/ru/knowledge-base/technology-innovation/big-data-bolshie-dannye. – Дата доступа: 26.08.2020.
2. **Хрисанфова, Е.** Облачные, туманные и граничные вычисления: отличия и перспективы развития технологий. Rusbase [Электронный ресурс]. – Режим доступа: https://rb.ru/story/edge-computing/. – Дата доступа: 03.09.2020.
3. Большие данные (Big Data). TAdvisor. [Электронный ресурс]. – Режим доступа: https://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_(Big_Data)#.D0.A1.D0.B0.D0.BC.D0.BE.D0.B5_.D0.BF.D1.80.D0.BE.D1.81.D1.82.D0.BE.D0.B5_.D0.BE.D0.BF.D1.80.D0.B5.D0.B4.D0.B5.D0.BB.D0.B5.D0.BD.D0.B8.D0.B5. – Дата доступа: 08.09.2020.
4. **Acharya, B., Jena, A. K., Chatterjee, J. M., Kumar, R., Le, D. N.** NoSQL Database Classification: New Era of Databases for Big Data // International Journal of Knowledge-Based Organizations (IJKBO). – 2019. – № 9 (1). – P. 50–65. [Electronic resource]. – Access mode: https://econpapers.repec.org/ article/iggjkbo00/v_3a9_3ay_3a2019_3ai_3a1_3ap_3a50-65.htm. – Access date: 14.08.2020.

The article gives a definition of the term «big data», identifies the features of big data, describes the methods and technologies of it, reveals the differences between traditional methods of data processing and big data technologies, determines the importance of using big data by enterprises.

Key words: *big data, methods of analysis big data.*

Получено 05.10.2020.