Свириденкова М.А., Свириденков К.И. Тенденции развития Big Data // Международный журнал информационных технологий и энергоэффективности. – 2020. – Т.5 № 1(15) с. 23-29

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.6

ТЕНДЕНЦИИ РАЗВИТИЯ BIG DATA

Свириденкова М.А., Свириденков К.И.
Филиал ФГБУ ВО «НИУ «МЭИ» в г. Смоленске, Смоленск, Россия (214013, г. Смоленск, Энергетический проезд,1), e-mail: sviridenkova-m-a@mail.ru

Ключевой идеей в данном исследовании выступает определение тенденций развития технологий Big Data. Представлена история возникновения Big Data и формирования понятия «большие данные». Выделено пять основных тенденций развития больших данных. Обоснованы положительные стороны использования в организациях технологий Big Data в процессах принятия решений и выработке стратегических направлений развития организации.

Ключевые слова: большие данные, технологии, информация, принятие решений.

BIG DATA DEVELOPMENT TRENDS

Sviridenkova M.A., Sviridenkov K.I.
Smolensk Branch of the National Research University "Moscow Power Engineering Institute", Smolensk, Russia (214013, Smolensk, Energetichesky proezd, 1), e-mail: sviridenkova-m-a@mail.ru

The key idea in this study is the identification of trends in the development of Big Data technologies. The history of the emergence of Big Data and the formation of the concept of "big data" is presented. Five main trends in the development of big data are highlighted. The positive aspects of the use of Big Data technologies in organizations in decision-making processes and the development of strategic directions for the development of organizations are substantiated.

Keywords: big data, technology, information, decision making.

Сегодня BIG DATA являются ключевым фактором развития информационных технологий. В эпоху информационных технологий и особенно цифровизации всех сфер человеческой деятельности по каждому пользователю Интернета накапливается большое количество информации. Это во многом определяет развитие направления BIG DATA. Термин Big Data нельзя считать синонимом информации или анализа информации. Информация чаще всего может быть представлена в неструктурированном виде, например, это могут быть видеозаписи, машинный код, изображения, текстовые документы и не только. К тому же эти данные могут быть разбросаны по различным хранилищам, которые могут находиться даже на разных континентах. Термин «Большие данные» означает большой объем накопленной информации, а также технологии хранения, вычисления, сервисные услуги.

Значительная часть информации создается не столько человеческими усилиями, сколько с использованием автоматизированных устройств. Причем такие устройства взаимодействуют и друг с другом, и с другими сетями данных, например, сенсоры и интеллектуальные устройства. По прогнозам исследователей, объем данных в мире будет

ежегодно удваиваться. Во много раз вырастет количество виртуальных и физических серверов, в частности, за счет расширения и создания новых data-центров. Поэтому растет потребность в эффективной обработке, анализе больших данных в целях их использования в процессах принятия решений. В связи с этим интересным представляется определение тенденций развития технологий Big Data.

Впервые термин «большие данные» упоминается в 2005 году. Однако до сих пор не существует его устоявшегося определения, которое было бы принято в научном и профессиональном сообществе [1]. Возможно, это связано с гибкостью понятия, за которым находится постоянно изменяющаяся сфера разработки новых средств работы с данными. А может быть причиной этой неопределенности является огромное количество различных областей, в которых «большие данные» могут быть применены: экономика, социология, медицина, ритейл, Интернет вещей, промышленность и многие другие [2].

Феномен «Большие данные» сопровождается тремя основными характеристиками:

- volume – большой объем;
- velocity – большая скорость поступления новых данных;
- variety – разнообразие и неструктурированность.

Область понимания BIG DATA постоянно изменяется, а точнее расширяется. Например, к трем указанным выше характеристикам добавляются другие, тоже начинающиеся на “V” (например, veracity, value). В результате термин обрастает новыми признаками и эволюционирует. Отсюда можно сделать вывод, что спрос на технологию «больших данных» растет [2].

В результате использования технологии Big Data предполагается решение трех типов задач: 1) хранение и управление большими объемами данных в сотни терабайт ; 2) организация неструктурированной информации; 3) анализ Big Data. Под большими объемами данных понимаются данные в сотни терабайт, которые неэффективно обрабатывать с использованием реляционных баз данных. Неструктурированную информацию тоже сложно обрабатывать с использованием таких баз данных и, соответственно, сложно формировать аналитические отчеты и разрабатывать прогностические модели развития исследуемых явлений.

В сфере Big Data существует много направлений. Среди них можно выделить два основных: Big Data engineering, Big Data Analytics (Scientist). Big Data engineering занимается сбором, хранением, преобразованием данных к виду, которого требуют приложения для корректной работы. Big Data Analytics — это следующая часть, в которой возможно использование объемных данных из уже созданных систем. Анализ состоит из расширенных вычислений по данным, где происходит прогнозирование, выявление закономерностей, тенденций и построение классификаций.

Основной движущей силой технологии является постоянно растущий объем производимых и потребляемых данных. Аналитики предсказывают, что к 2025 году этот объем достигнет 163 Збайт (1 Збайт = 1 триллион Гбайт) [3]. Технология «больших данных» стремительно развивается, расширяются и направления их использования. Сегодня спрос на «большие данные» формируется как в государственном, так и в частном секторе. Привлекательным для предприятий является возможность использовать технологию в принятии решений. Поэтому перспективы развития технологии BIG DATA огромные. Они обусловлены увеличением количества данных, поступающих на обработку, а также ужесточением требований к скорости их обработки. Пока невозможно понять, какая страна или корпорация займет в будущем лидирующее положение в этой области. Но уже очевидно, что тот, кто овладеет методами хранения и анализа огромных массивов данных, получит мощнейший инструмент для оптимизации принятия решений и выбора стратегии развития.

Технология все еще находится на раннем этапе своего внедрения, поэтому наблюдать итоги ее реализации в конкретных проектах затруднительно. Сегодня разрыв между «большими данными» и их внедрением в реальных сферах жизни продолжает убывать.

Таким образом, уже можно наблюдать их использование и выделить некоторые тенденции развития.

Первая тенденция развития «больших данных» – разработка аппаратных и программных средств, осуществляющих работу с информацией, распределенной на многих компьютерах. Понятие «большие данные» начинает применяться тогда, когда одного современного компьютера становится недостаточно для обработки и анализа массива информации. То есть, когда речь идет о распределенных системах. Существующие классические СУБД плохо масштабируются в таких случаях, поэтому необходимо создавать новые удобные инструменты. К таким средствам относятся, например, фреймворк Hadoop и СУБД Apache Hive. Но развиваются и новые системы, улучшающие скорость работы с данными – например, отечественное модульное хранилище данных Exarch.

Сегодня основополагающая технология – Hadoop, проект Apache Software Foundation, используемый для реализации поисковых и контекстных механизмов высоконагруженных веб-сайтов [4]. Hadoop свободно распространяет набор утилит, библиотек и программный каркас для разработки и выполнения распределенных программ, работающих на кластерах из множества узлов. Hadoop является основной платформой развития российского рынка, способствующей повышению отечественной конкуренции. В таблице 1 представлены преимущества платформы Hadoop.

Таблица 1 – Преимущества платформы Hadoop

Преимущество	Краткое описание
Снижение времени на обработку данных	Существенное уменьшение времени обработки данных на кластере
Снижение стоимости оборудования	Сокращение затрат на оборудование, требуемое для хранения и обработки данных, в десятки раз.
Повышение отказоустойчивости. Технология позволяет построить отказоустойчивое решение	Выход узла, влияет лишь на производительность системы, но не на ее работу.
Линейная масштабируемость	Наращивание производительности.
Работа с неструктурированными данными	Обработка любых файлов, в том числе неструктурированных. Повышение полезности информации.

Вторая тенденция – разработка инструментов, оперирующих большими массивами данных. Несмотря на наличие некоторого парка инструментов, технология Big Data не так давно перешла с этапа академических исследований к применению в реальности. Уже разработаны и успешно используются инструменты, позволяющие оперировать данными огромных объемов. Однако их всеобщее внедрение еще не наступило. Применение Big Data носит избирательный, хоть и масштабный характер. Например, в Китае работает система «социального кредита», которая определяет рейтинг граждан по их действиям. Эти действия отслеживаются системами массового видеонаблюдения, а поступающая информация обрабатывается с помощью технологий «больших данных». В результате мы наблюдаем еще одну тенденцию: основным ресурсом технологии становится объем обрабатываемых данных в отличие от имевшегося ранее, на этапе академических исследований, - трудовых ресурсов профессионалов. В качестве примера можно привести аналогию с добычей нефти: богаче будет та страна, в которой находятся огромные запасы нефти, а не та, в которой будут сконцентрированы лучшие профессионалы нефтедобычи. При этом предполагается и прогнозируется, что эффективные методы работы с колоссальными массивами информации естественным образом возникнут там, где этой информации наибольшее количество, так как ее необходимо уметь эффективно обрабатывать.

Третья тенденция – централизация хранения данных. На практике могут быть использованы централизованные, децентрализованные и смешанные хранилища данных. В рамках централизованного хранилища справочная информация извлекается из различных источников, систематизируется, дополняется, записывается в эталонное хранилище. На текущий момент времени такая организация хранения данных получила наибольшее распространения, в основном, из-за простоты и высокой скорости доступа к анализируемой информации. Принцип централизованного хранения реализуют системы, указанные в таблице 2.

Таблица 2 – Продукты, использующие принцип централизованного хранения данных

Продукты	Применение
1. IBM Client Information Integration Solution (IBM CIIS)	Хранилище данных, использующее пакетный режим и обработку данных в режиме реального времени. Управление осуществляется через графическую оболочку, что сопровождается снижением требований к квалификации сотрудников, обеспечивающих поддержку нормативно-справочной информации. На сегодняшний день IBM CIIS используют преимущественно банки и страховые организации.
2. Oracle Castomer Data Hub (Oracle CDH)	Продукт может использоваться для управления реестрами клиентов, сотрудников, населения как отдельных регионов, так и страны в целом. На сегодняшний момент времени продукт используют в телекоммуникационных и высокотехнологичных компаний.
3. SAP Master Data Management (SAP MDM)	Данный продукт служит вспомогательным инструментом для управления коллаборативным и гибким бизнесом. Используемая технология – MDM (Управление основными мастер-данными). В его основе лежит обеспечение тесных свизей с деловыми партнерами. При этом предоставляется возможность централизованного ведения справочников и классификаторов, поддерживаются процессы изменения информации в справочниках, обеспечивается распределенный доступ пользователей к этим данным, а также распространение согласованных данных между всеми приложениями предприятия. Разработчик продукта немецкий ИТ-гигант SAP AG

При использовании децентрализованного хранилища справочной информации создается виртуальная база данных, в случае обращения к которой идет запрос данных к тем системам, где они хранятся непосредственно. В результате информация по одному клиенту будет собрана из фрагментов, содержащихся в разнородных базах данных, но в виртуальной базе данных она будет представлять собой одну запись. Продукты, реализующие децентрализованное хранение данных, MetaBase и MetaMatrix Server, разработчик - компания MetaMatrix.

Находят применение и смешанные решения, использующие одновременно принципы централизованного и децентрализованного хранения данных. Они представлены ниже, в таблице 3.

Таблица 3 – Продукты, использующие принцип децентрализованного хранения данных

Продукты	Применение
1. Initiate Systems Identity Hub	Разработка компании Initiate Systems. Ранее было использовано в централизованной архитектуре.. Особенно выделяется система сверки данных. На сегодняшний день продукт широко используется в фармацевтических компаниях и организациях здравоохранения, а также благодаря динамичному развитию стремится занять и другие ниши рынка управления нормативно-справочной информацией.
2. DWL Customer	Разработка компании DWL. Представляет собой гибридное хранилище данных, которое обладает лучшей среди представленных в статье систем поддержки сервисно-ориентированной архитектуры (Service-Oriented Architecture, SOA). Продукт внедрен в финансовых организациях, но предполагает внедрение решения для сферы телекоммуникаций и здравоохранения;
3. Siebel Universal Customer View (Siebel UCM)	Аналогично DWL Customer, обладает развитой поддержкой архитектуры SOA; в составе продукта насчитывается около 140 Web-сервисов.
4. Siperian Master Reference Manager (Siperian MRM)	Продукт представляет собой систему с гибкой моделью хранения, включающую: 1) эталонную базу данных, по которой можно идентифицировать сущности и ключи соответствия между используемыми источниками информации; 2) другие источники, включающие дополнительные (например, операционные) характеристики сущности. Характеристики из других источников загружаются в эталонную базу данных.

Таким образом, среди централизованных, децентрализованных и смешанных хранилищ данных наибольшее распространение получили первые из перечисленных. К ним можно отнести облачные хранилища и дата-центры. С целью практического использования больших объемов собранной информации предварительно ее необходимо систематизировать. Тенденция к централизации и дисциплинированному управлению данными, несомненно, очень важна при принятии решений.

Качество обработки информации целесообразно оценивать путем каких-либо показателей. Как известно, показатели могут носить индивидуальный, групповой, комплексный характер (или интегральный). На сегодняшний день компания IDC в рамках исследований «Глобальная цифровизация от периферии к центру» («The Digitization of the World – from Edge to Core) разработала один из таких показателей. Это индекс DATCON (DATA readliness CONdition) - «уровень готовности к работе с данными», который позволяет руководителям определить, на каком уровне в организации находятся управление данными, их использование и монетизация [5]. Это интегральный показатель. Его значение находится в диапазоне 1 – 5 (1 – критичный, 5 – оптимизированный). При расчете индекса учитываются следующие факторы: темпы изменения объемов данных, ценность данных, уровень информационной безопасности организации, размеры инвестиций, зрелость и качество управления, наличие у персонала организации необходимых навыков по обработке больших массивов данных, степень вовлеченности топ-менеджеров организации в проекты, связанные с данными. В рамках исследований условий работы с постоянно увеличивающимся объемом

данных, проведенных компанией IDC в четырех областях (производство, финансовый сектор, здравоохранение, СМИ и индустрия развлечений), лидирует производство. В целом индекс DATCON может быть использован в процессах принятия решений в различных областях и особенно в производственной деятельности организаций. Предполагается, что это позволит выработать наиболее эффективное стратегическое решение.

Четвертая тенденция, связанная с получением данных – это развитие IoT (интернет вещей) технологий, которые обеспечивают автоматизированный сбор данных на самых различных типах устройств. На примере автомобилестроения можно видеть, что данная технология уже успешно используется. Данные в реальном времени, поступающие от автомобилей, позволяют оказывать потребителям более грамотный сервис: начиная от исправления поломки до поиска угнанного транспортного средства. Все больше компаний внедряют IoT-технологии в свои продукты, что приводит к необходимости овладения методами сбора больших массивов информации. В целом это выгодно для производителя, так как позволяет быстрее и полнее получать информацию о работе своих продуктов для их дальнейшего развития. Рынок интернета вещей продолжает расти, и сейчас сложно найти сферу человеческой жизнедеятельность, где он бы не использовался (рисунок 1).

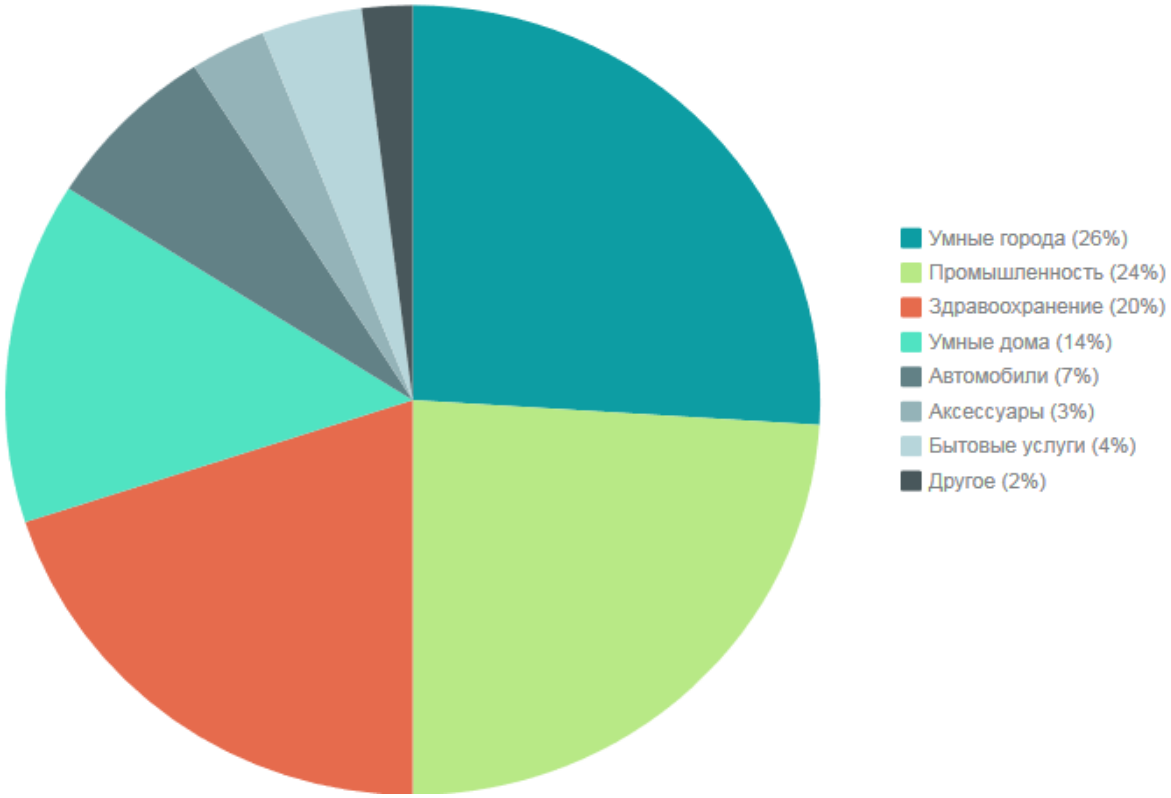

Рисунок 1 – Использование IoT в различных сферах в мире

Пятая тенденция связана с анализом полученных данных. В этой области все большую роль играет искусственный интеллект [6]. Современные алгоритмы машинного обучения на данный момент способны анализировать данные лучше человека во многих областях. Например, сейчас искусственный интеллект по образцам крови может лучше предсказать заболевание пациента, чем самый опытный врач-гематолог. Тот факт, что машина допускает меньше ошибок и имеет лучшую способность к предсказанию, позволяет предположить, что постепенно человеческий ресурс во многих областях будет заменен компьютерным. И это на самом деле является очередной тенденцией развития работы с данными. Искусственный интеллект активно используется многими технологическими компаниями нашего времени именно потому, что может обработать огромный массив информации и выдать адекватный результат. Например, виртуальный голосовой помощник «Алиса», созданный компанией

«Яндекс», хранит огромную базу информации, помогающую отвечать на запросы пользователей. Компьютер можно обучить работать над специализированной задачей быстрее и лучше, чем человека. Возможно, в будущем человек будет принимать только самые важные решения, доверяя какие-то рутинные задачи роботам и алгоритмам.

В России развитие и использование в практической деятельности технологий Big Data определяются требованиями цифровизации экономики, реализуемыми программой «Цифровая экономика Российской Федерации». Указом Президента РФ от 9 мая 2017 г. № 203 утверждена государственная программа «О Стратегии развития информационного общества в Российской Федерации на 2017 - 2030 годы», в рамках которой планируется развивать технологии Big Data [7].

Таким образом, в данной работе рассмотрены некоторые аспекты рынка больших данных, выявлены положительные стороны и тенденции развития данного направления, а также обоснована целесообразность использования Big Data на российском рынке.

Список литературы:

1. Большие данные // Википедия, 2017. [Электронный ресурс]. Режим доступа: http://ru.wikipedia.org/?oldid=87934960/
2. Васильев В.И. Обзор технологий для работы с Big Data // Молодой ученый. – 2020. - №9. – с. 13 – 14. -Режим доступа : https://moluch.ru/archive/299/67818/
3. Резванов А. К 2025 году общий объем данных в мире достигнет 163 зеттабайт. [Электронный ресурс]. Режим доступа: http://www.macster.ru/news/170412-k-2025-godu-obshchiy-obem-dannykh-v-mi.
4. Аналитический обзор рынка Big Data. [Электронный ресурс] – Режим доступа: https://www.sostav.ru/publication/big-data-razmer-ne-imeet-znachenie-31028.html
5. The Digitization of the World From Edge to Core. David Reinsel – John Gantz – John Rydning
6. Митрович С. Рынок «больших данных» и их инструментов: тенденции и перспективы в России // МИР (Модернизация. Инновации. Развитие). 2018. Т. 9. № 1. С. 74–85. DOI: 10.18184/2079–4665.2018.9.1.74–85
7. Указ Президента Российской Федерации от 09.05.2017 № 203 "О Стратегии развития информационного общества в Российской Федерации на 2017 - 2030 годы" [Электронный ресурс] – Режим доступа: http://publication.pravo.gov.ru/Document/View/0001201705100002

References

1. Big data // Wikipedia, 2017. [Electronic resource]. Access Mode: http://ru.wikipedia.org/?oldid=87934960/
2. Vasiliev V.I. Overview of technologies for working with Big Data // Young scientist. - 2020. - No. 9. - with. 13 - 14. -Access mode: https://moluch.ru/archive/299/67818/
3. Rezvanov A. By 2025, the total amount of data in the world will reach 163 zettabytes. [Electronic resource]. Access mode: http://www.macster.ru/news/170412-k-2025-godu-obshchiy-obem-dannykh-v-mi.
4. Analytical review of the Big Data market. [Electronic resource] - Access mode: https://www.sostav.ru/publication/big-data-razmer-ne-imeet-znachenie-31028.html
5. The Digitization of the World From Edge to Core. David Reinsel - John Gantz - John Rydning
6. Mitrovich S. Market of "big data" and their tools: trends and prospects in Russia // MIR (Modernization. Innovations. Development). 2018.Vol. 9. No. 1. P. 74–85. DOI: 10.18184 / 2079–4665.2018.9.1.74–85
7. Decree of the President of the Russian Federation of 05.09.2017 No. 203 "On the Strategy for the Development of the Information Society in the Russian Federation for 2017 - 2030" [Electronic resource] - Access mode: http://publication.pravo.gov.ru/Document/View / 0001201705100002.