УДК 004.041

# ТЕХНОЛОГИИ БОЛЬШИХ ДАННЫХ И ПЕРСПЕКТИВЫ ИХ ВНЕДРЕНИЯ В РЕСПУБЛИКЕ БЕЛАРУСЬ

Белодед Н.И., Иванова И.А., Качанова Е.А.

*Академия управления при Президенте Республики Беларусь, г.Минск, Республика Беларусь, nbeloded@gmail.com, ms.ira.99@mail.ru, katya_16_08@mail.ru*

***Аннотация.*** *Статья посвящена такому понятию, как технологии больших данных. Рассмотрены теоретические аспекты больших данных, проанализированы методы, функции, принципы и задачи данной технологии, а также перспективы внедрения данной методики в различных сферах жизни белорусского общества.*

С развитием технологий в мире значительно возросло количество данных. Так, в мире в середине 2017 г. насчитывалось 3885,5 млн пользователей сети Интернет. Рейтинг стран по количеству интернет-пользователей представлен на рисунке 1.


Рисунок 1 – Рейтинг стран по количеству интернет-пользователей

На рост потребителей (и создателей) интернет-контента значительно влияет удешевление и как следствие повышение доступности интернета для широкого круга пользователей. Так, по мнению специалистов IHS Markit, количество пользователей смартфонов в 2014 г. Было 1,57 млрд чел., а по прогнозам к 2020 г. этот показатель достиг 2,87 млрд чел. Как следствие, растет генерирование и накопление значительных объемов информации, обеспечение возможности обмена и свободного доступа к ней.

Вместе с этим у человечества возникает потребность в определенного рода инструментах, способных качественно и оперативно извлекать пользу из больших, а порой и гигантских массивов информации. Возросла необходимость применения специальных технологий, с помощью которых государственные структуры и представители бизнеса могли бы с минимальными затратами оптимизировать различные процессы, а конечные потребители получать более качественные услуги. Самым эффективным подходом к решению вышеуказанных задач являются технологии больших данных или Big Data.

**Большие данные** (англ. Big Data) – «…серия подходов, инструментов и методов обработки структурированных и неструктурированных данных огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам вычислительной сети, сформировавшихся в конце 2000-х годов, альтернативных традиционным системам управления базами данных». [1]

Таким образом, к Big Data можно отнести не только сами данные, но и технологии их обработки и применения, а также методы поиска необходимой информации в больших массивах. Проблема больших данных по-прежнему остается открытой и жизненно важной в рамках любых систем, десятилетиями накапливающих самую разнообразную информацию.

Первое упоминание термина «большие данные» в литературе было сделано в 2008 году в журнале «Nature» Клиффордом Линчем. Это была статья на тему перспективного развития науки посредством технологий взаимодействия с большим количеством данных. Ранее данный термин рассматривался лишь с точки зрения научного анализа, однако с релизом еще нескольких статей пресса стала широко использовать понятие Big Data – и продолжает применять его в настоящее время.

В 2010 году для решения возрастающей проблемы больших данных, были выпущены программные компоненты, направленные в первую очередь на то, чтобы сократить степень риска от применения крупных массивов информации.

К 2011 году многие влиятельные компании, такие как Microsoft, Oracle, EMC и IBM всерьез заинтересовались технологией Big Data и начали успешно применять данные разработки в рамках собственных стратегий развития. А американская консалтинговая компания Gartner, специализирующаяся на исследованиях рынка информационных технологий, в свою очередь, отметила большие данные как тренд номер два (после виртуализации) в сфере IT.

По данным компании IBS, к 2003 году мировое сообщество аккумулировало 5 эксабайтов данных (1 ЭБ = 1 млрд гигабайтов). К 2008 году этот объем возрос до 0,18 зеттабайта (1 ЗБ = 1024 эксабайта), к 2011 году – до 1,76 зеттабайта, к 2013 году – до 4,4 зеттабайта. В мае 2015 года глобальное количество данных перешло за черту 6,5 зеттабайта. К 2020 году, по подсчетам ученых, в мире уже сформировано 40-44 зеттабайтов информации.

А к 2025 году по прогнозам ученых это число преумножится в 10 раз, говорится в докладе «The Data Age 2025», который был подготовлен аналитиками компании IDC. В докладе отмечается, что большую часть данных генерировать будут сами предприятия, а не обычные потребители. [2]

В вышеуказанном исследовании также делается предположение, что данные станут жизненно-необходимым ресурсом, а безопасность – критически важным фундаментом в жизни. Также авторы отмечают, что технология существенно повлияет на экономическую ситуацию, а что касается обычного пользователя, то время взаимодействия с подключёнными устройствами значительно возрастет и будет составлять около 4800 раз в день. Таким образом, проблема больших данных будет оставаться актуальной в ближайшие десятки, а может и сотни лет. Для ее решения необходимо четко сформулировать механизмы обработки и алгоритмы взаимодействия с большими информационными массивами.

В первую очередь, необходимо выявить источники накопления и появления больших данных. К ним можно отнести**:**

Интернет – социальные сети, блоги, СМИ, форумы, сайты, интернет вещей (IoT).

Корпоративные данные – транзакционная деловая информация, архивы, базы данных.

Показания устройств – датчиков, сенсоров, приборов, а также метеорологические данные, данные сотовой связи и т. д.

Таким образом, многие привычные человеку элементы бытовой среды сегодня оборудованы устройствами для сбора данных – сенсорами, измерителями, которые в

результате получения и обработки различного рода данных, принимают за нас решения и тем самым помогают нам жить.

Необходимо также отметить, что как только человечество переступило рубеж цифровой эпохи, сбор так называемых «старых данных» стал осуществляться быстрее, что сделало возможным способность их накапливать. Это в свою очередь привело к появлению больших массивы данных, из которых можно извлекать качественно иную информацию, причем уже не просто для контроля тех или иных действий в любой сфере жизнедеятельности, например, сколько раз пациент прошел флюорографию, а для анализа тенденций, как в рамках отдельной страны, так и мира в целом.

**Итак, к основным методам анализа и обработки данных можно отнести следующие:**

**Методы класса или глубинный анализ (Data Mining)** подразумевают под собой широкий набор методов для извлечения ранее неизвестных, нетривиальных, практически полезных знаний из так называемых «сырых», необходимых для принятия решений. Такие методы, в частности, включают изучение правил ассоциации, классификацию (категоризацию), кластерный анализ, регрессионный анализ, обнаружение, анализ отклонений и т.д.

**Краудсорсинг.** Этот метод позволяет одновременно получать данные из нескольких источников, и их количество практически не ограничено.

**Статистический анализ или A/B-тестирование.** Метод маркетингового исследования, основан на том, что среди общего объема данных выбирается набор элементов управления, который поочередно сравнивается с другими аналогичными наборами, в которых один из элементов был изменен. Выполнение таких тестов помогает определить колебания параметров, которые оказывают наибольшее влияние на контрольную популяцию. С большими объемами данных можно выполнять большое количество итераций, каждая из которых обеспечивает наиболее надежный результат.

**Прогнозная аналитика.** Методология основана на том, что специалисты в этой области стараются заранее прогнозировать и планировать, как поведет себя контролируемый объект, чтобы принять наиболее выгодное решение в конкретной ситуации.

**Машинное обучение (искусственный интеллект)** базируется на эмпирическом анализе информации и создании алгоритмов самообучения систем.

**Сетевой анализ** является наиболее распространенным методом исследования социальных сетей, состоящий в том, что после получения статистических данных анализируются узлы, сформированные в сети, то есть взаимодействия между отдельными пользователями и их сообществами.

**Искусственные нейронные сети**, сетевой анализ, оптимизация, в том числе генетические алгоритмы (англ. genetic algorithm − алгоритмы эвристического поиска для решения задач оптимизации и моделирования путем случайного отбора, объединения и вариации требуемых параметров с использованием механизмов, аналогичных естественному отбору).

**Имитационное моделирование** (англ. simulation) − метод, который базируется на построении моделей, которые призваны описывать процессы так, как они проходили бы в действительности. Имитационное моделирование можно рассматривать как разновидность экспериментальных исследований.

**Пространственный анализ** (англ. spatial analysis) представляет собой целый класс совокупных методов, манипулирующих топологической, геометрической и географической информацией.

**Визуализация аналитических данных** − графическое представление информации в виде всевозможных рисунков, диаграмм, графиков, в том числе с применением интерактивных возможностей и анимации как с целью получения результатов, так и для применения в в качестве данных для последующего анализа. Значительная стадия анализа

больших массивов данных, поскольку помогает представить самые важные, порой и не самые явные результаты анализа в наиболее удобном для восприятия виде.

Для корректного функционирования система больших данных должна быть основана на определенных **принципах**:

1. **Горизонтальная масштабируемость**. На том основании, что данных может быть бесконечно много, любая система, связанная с обработкой больших данных, вне зависимости от сферы ее применения, должна быть расширяемой.

2. **Отказоустойчивость**. Исходя из предыдущего принцип горизонтальной масштабируемости, машин в кластере может быть много. Например, Hadoop-кластер Yahoo сосредотачивает более 42000 машин. Это непременно означает, что часть этих машин будет с большой вероятностью выходить из строя. Методы работы с большими данными должны учитывать возможность подобных сбоев и преодолевать их без каких-либо серьезных последствий.

3. **Локальность данных.** В крупных распределённых системах данные сосредоточены внутри большого количества вычислительных машин. Притом данные физически находятся на одном сервере, а обрабатываются на другом, и как следствие расходы на передачу данных могут превысить расходы на непосредственную обработку. Поэтому при проектировании BigData-решений рекомендуется опираться на принцип локальности данных. Он подразумевает обработку данных на той же машине, на которой осуществляется их хранение.

Все современные средства работы с большими данными так или иначе следуют этим трём принципам. Для того, чтобы им следовать – необходимо руководствоваться определенными методами, способами и парадигмами разработки средств разработки данных.

**Функции и задачи больших данных**

Когда говорят о Big Data, упоминают правило VVV – три признака или **свойства**, которыми большие данные должны обладать:

Volume – объем (данные должны измеряться по величине физического объема документов).

Velocity – скорость обновления, то есть данные должны регулярно обновляться, что в дальнейшем потребует их постоянной обработки.

Variety – разнообразие данных различных форматов, как неструктурированные, так и частично структурированные.

**Перспективы и тенденции развития Big data**

Уже начиная с 2017 года, когда большие данные не были чем-то новым и неизвестным, их важность и значение использования не только не уменьшилась, а еще более возросла. Исходя из этого эксперты делают ставку на то, что анализ больших объемов данных станет доступным не только для крупных организаций, но и для представителей малого и среднего бизнеса. Данный подход можно реализовать с помощью следующих компонентов:

**1. Облачные хранилища.** С их помощью обработка и хранение данных становятся намного быстрее и экономичнее – если же сравнивать с расходами на содержание собственного дата-центра и с возможным расширением персонала, то аренда облака представляется наиболее дешевой альтернативой.

**2. Blockchain.** С помощью этой технологии вы сможете ускорить и упростить многочисленные интернет-транзакции, в том числе и международные. Еще одним преимуществом Блокчейна является то, что благодаря ему можно снизить затраты на проведение транзакций.

**3. Dark Data.** Использование так называемых «темных данных», представляющие собой всю не цифрованную информацию об организации, которая не играет ключевой роли в ее прямом использовании, но она может стать причиной для перехода на новый формат хранения каких-либо сведений.

**4. Самообслуживание и снижение цен.** Начиная с 2017 года планировалось внедрение «платформы самообслуживания» − это бесплатные площадки, на которых представители малого и среднего бизнеса могут самостоятельно оценить и систематизировать хранящиеся у них данные.

**5. Искусственный интеллект и Deep Learning.** Технология обучения машинного интеллекта, которая имитирует структуру и работу человеческого мозга, лучше всего подходит для обработки большого количества постоянно меняющейся информации. В этом случае машина сделает все то, что должен был сделать и человек, но вероятность ошибки при этом значительно уменьшится.

Сразу же после того, как началось активное внедрение технологий Big Data на рынок и в современную жизнь, их стали использовать всемирно известные компании-гиганты, которые имеют клиентов практически в каждой точке земного шара..

К таким компаниям относятся: Google и Facebook, IBM., а также финансовые структуры такие, как VISA, Master Card, и Bank of America.

Например, IBM применяет методы больших данных к проводимым денежным транзакциям. С их помощью было выявлено на 15% больше мошеннических транзакций, что увеличило объем защищенных средств на 60%. Кроме того были решены проблемы с ложными срабатываниями системы − их количество сократилось более, чем в два раза.

Компания VISA так же, как и IBM, использовала Big Data для того, чтобы отслеживать мошеннические попытки произвести ту или иную операцию. Именно благодаря этому ежегодно они спасают от утечки более 2 млрд долларов США.

Например, Министерству труда Германии удалось сократить свои расходы на 10 млрд евро, внедрив систему больших данных в работу по выдаче пособий по безработице. В то же время выяснилось, что пятая часть граждан безосновательно получает данные пособия.

Также Big Data не обошли стороной и игровую индустрию. Например, разработчики World of Tanks провели исследование информации обо всех игроках и сравнили имеющиеся показатели их активности. С помощью этого исследования им удалось спрогнозировать возможный будущий отток игроков − исходя из сделанных предположений, представители организации смогли более эффективно взаимодействовать с пользователями.

HSBC, Starbucks, Nasdaq, Coca-Cola, и AT&T можно отнести к числу известных организаций, которые используют большие данные.

Социальные сети, на сегодняшний день, являются настоящим царством огромнейших массивов данных, которые предоставляют ценную информацию о каждом потенциальном клиенте. Например, пользователи крупнейшей социальной сети в мире Facebook (более 1,32 млрд чел.) 3 млрд раз в день нажимают кнопку «Нравится»; каждый час 4,5 млн человек получают приглашение на мероприятие. Каждый пользователь оставляет за собой цифровой след, который помогает компаниям изучить не только предпочтения пользователей, но и их поведение.

Обрабатывая большие объемы данных, можно решить многие глобальные проблемы в таких областях, как бизнес, веб-аналитика, медицина, образование и многое другое.

Что касается Республики Беларусь, к сожалению, не все владельцы и руководители крупнейших предприятий Беларуси осознали, насколько важно эффективно обрабатывать большие объемы данных, как они могут помочь предприятию работать, сохранить и приумножить деньги их хозяев. Например, согласно статистике сайта belmarket.by, тех, кто осознал необходимость внедрения BI (бизнес-анализа), меньше половины даже в секторе крупных предприятий. [3]

Для того, чтобы улучшить условия внедрения технологий больших данных в белорусскую экономику, а также реализовать необходимость выработки новых подходов к созданию научно-технической основы экономики, которые определяют будущее динамичное поступательное движение Беларуси по инновационному пути, была разработана

долгосрочная стратегия формирования и развития модели белорусской экономики, основанной на интеллекте, − Стратегия «Наука и технологии: 2018–2040».

К приоритетным технико-технологическими направлениям «Новой Индустрии 2040» относятся: создание общенациональной сети больших данных, программного обеспечения и суперкомпьютеров для обеспечения всестороннего взаимодействия между предприятиями реального сектора, а также системами идентификации и отслеживания товаров. [4]

Таким образом, в Республике Беларусь использование технологии Big Data обсуждается на государственном уровне и первая отрасль, которая применит обработку и аналитику больших данных у себя, станет здравоохранение. Об этом заявил на открытии международной конференции BIG DATA − 2017 в Минске первый заместитель министра связи и информатизации Дмитрий Шедко. [5]

3 мая 2017 года в Национальной библиотеке Беларуси состоялась 3-я Международная научно-практическая конференция «BIG DATA and Advanced Analytics (big data и анализ высокого уровня)» (BIG DATA-2017). Она объединила более 300 экспертов и ученых в области обработки и анализа больших объемов данных, разработок и внедрения новых технологий по соответствующему направлению.

По мнению Шедко, следующей отраслью, которая применит на практике большие данные после здравоохранения будет реальный сектор экономики, поскольку от внедрения технологии big data будет зависеть конкурентоспособность предприятий и их эффективность. «До 2020 года во всех перечисленных мною сферах в разной степени эти системы будут внедрены. Иначе мы будем тотально не эффективны», − уверенно заявил Шедко.

Что касается именно сферы образования, то, по словам заместителя министра образования Республики Беларусь, кандидата педагогических наук, доцента Ирины Старовойтовой, «цифровая трансформация обучения в ближайшем будущем будет реализовываться на основе анализа большого потока информации, поступающего в информационную среду образования от преподавателей, студентов, а также администрации учреждений образования. На сегодняшний день, генерируется большое количество различных типов данных, обработку которых трудно осуществить традиционными математическими методами. Необходимы но-вые технологии для хранения и обработки этих данных.

В Академии управления при Президенте Республики Беларусь на кафедре управления информационными ресурсами разрабатывается информационная система оценки компетенций профессорско-преподавательского состава, предусматривающая аналитическую обработку больших данных. В роли экспертов для оценки компетенций выступают: руководство Академии, сотрудники внешних организаций – партнеров, а также коллеги-сотрудники и студенты. Разрабатываемая информационная система позволяет в полной мере реализовать все возможные преимущества технологии Big Data и значительно улучшить образовательный процесс.

На примере можно рассмотреть компанию Деловая сеть, которая активно инвестирует в развитие единого высокоскоростного бесплатного городского Wi-Fi-пространства, а также в сбор и анализ Big Data (больших данных), генерируемых на базе функционирования собственной сетевой инфраструктуры компании. Так, например, статистика подключений пользователей к собственной сети Wi-Fi превышает 5,5 млн в год. До 2020 года "Деловая сеть" планирует увеличить этот показатель в несколько раз. На сегодняшний день, бесплатная публичная Wi-Fi-сеть компании − Wi-Fi.BN.BY − доступна на более чем 600 точках по городу Минску и в областных центрах.

Так, например, для белорусского бизнеса обработка больших данных облегчит решение таких вопросов, как привлечение клиента, сокращение стоимости конечного продукта, увеличение прибыли или уменьшение риска банкротства. Кроме того, технология аналитики big data способна собирать данные о самой компании, ее внутренней среде, то есть о ее бизнес-партнерах, конкурентах, а также анализировать данные посредством социальных

сетей. Следует отметить, что торговые сети уже сейчас анализируют с помощью big data взаимозависимость продаж различных видов товара от его расположения на определенных полках, повышают лояльность текущих клиентов, оптимизируют интеграцию в цепи поставок, тем самым значительно упрощая планирование и увеличивая шансы проекта на востребованность.

Таким образом, резюмируя все вышесказанное, на сегодняшний день Big Data помогают решать различного рода задачи во многих сферах, среди них: ритейл, медицина, образование, финансы, промышленность, энергетика, туризм, экология, развлечения. Благодаря обработке и анализу большого массива данных, представители власти, бизнеса, науки, разработчики и другие заинтересованные лица улучшают качество товаров и услуг, развивают бизнес.

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ


1. Национальная библиотека им. Н. Э. Баумана. Big Data [Электронный ресурс]. – Минск, URL: https://ru.bmstu.wiki/Big_Data. – Дата доступа: 25.10.2019.
2. Data Age 2025: The Evolution of Data to Life-Critical Don’t Focus on Big Data; Focus on the Data That’s Big [Электронный ресурс] – Минск, URL: https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/workforce/Seagate-WP-DataAge2025-March-2017.pdf. – Дата доступа: 26.10.2019.
3. Белорусы и рынок. Дорасти до Big data – [Электронный ресурс]. – Минск, URL: http://www.belmarket.by/dorasti-do-big-data. – Дата доступа: 29.10.2019.
4. Стратегия «Наука и технологии:2018-2040» [Электронный ресурс]. – Минск, URL: http://nasb.gov.by/congress2/strategy_2018-2040.pdf. – Дата доступа: 29.10.2019.
5. Беларусь ждет взрывной интерес к большим данным: от здравоохранения до реального сектора – [Электронный ресурс] – Минск, URL: https://news.tut.by/economics/541969.html. – Дата доступа: 30.10.2018.