ПРОБЛЕМЫ И РЕШЕНИЕ В BIG DATA

И.Э.Абдирахимов
Каршинский инженерно-экономический институт, Карши, Узбекистан
E-mail: abdirahimov.ilhom@mail.ru

Аннотация. В этот статье представлены ключевые понятия по технологии Big Data: основные характеристики, методы, этапы перехода, сферы применения технологии. Произведен анализ мировых достижений в данной области, приведены примеры использования в работах зарубежных авторов. Проанализирован современный рынок использования технологии Big Data.
Ключевые слова: Big Data, Business Intelligence, технология, NewSQL, Hadoop, Data Mining, Методы.

PROBLEMS AND SOLUTION IN BIG DATA

I.E.Abdirahimov
Karshi Engineering-Economics Institute, Karshi, Uzbekistan
E-mail: abdirahimov.ilhom@mail.ru

Abstract. This article presents the key concepts for Big Data technology: the main characteristics, methods, stages of transition, the scope of technology. The analysis of world achievements in this area is made, examples of use in works of foreign authors are given. The modern market for the use of Big Data technology is analyzed.
Keywords: Big Data, Business Intelligence, technology, NewSQL, Hadoop, Data Mining, Metods.

Введение. Большие данные (англ. big data) – обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

В текущее время объемы информации растут по экспоненте. Для того чтобы быстрее реагировать на изменения рынка, получить конкурентные преимущества, повысить эффективность производства нужно получить, обработать и проанализировать огромное количество данных. Для работы с такими объемами информации инженеры были вынуждены модернизировать инструменты для работы над анализом всех данных. Сформировалось понятие BigData, которое было интересно лишь узкому кругу специалистов. Сейчас это слово на слуху у любого, кто интересуется

сферой информационных технологий. И это определение, а точнее направление развития ИТ, становится крайне популярным и стратегически важным в последнее время.

Методы исследования. Смешение и интеграция данных. Что это. Работа с big data часто связана со сбором разнородных данных из разных источников. Чтобы работать с этими данными, их нужно собрать воедино. Просто загрузить их в одну базу нельзя – разные источники могут выдавать данные в разных форматах и с разными параметрами. Тут и поможет смешение и интеграция данных – процесс приведения разнородной информации к единому виду.

Как это работает. Чтобы использовать данные из разных источников, используют следующие методы: приводят данные к единому формату; распознают текст с фотографий, конвертируют документы, переводят текст в цифры. Дополняют данные. Если есть два источника данных об одном объекте, информацию от первого источника дополняют данными от второго, чтобы получить более полную картину.

Отсеивают избыточные данные: если какой-то источник собирает лишнюю информацию, недоступную для анализа, ее удаляют.

Зачем и где применяют. Смешение и интеграция данных нужны, если есть несколько разных источников данных, и нужно анализировать эти данные в комплексе.

Например, ваш магазин торгует офлайн, через маркетплейсы и просто через интернет. Чтобы получить полную информацию о продажах и спросе, надо собрать множество данных: кассовые чеки, товарные остатки на складе, интернет-заказы, заказы через маркетплейс и так далее. Все эти данные поступают из разных мест и обычно имеют разный формат. Чтобы работать с ними, их нужно привести к единому виду.

Традиционные методы интеграции данных в основном основаны на процессе ETL извлечение, преобразование и загрузка. Данные получают из источников, очищают и загружают в хранилище. Специальные инструменты экосистемы больших данных от Hadoop до баз данных NoSQL также имеют собственный подход для извлечения, преобразования и загрузки данных.

После интеграции большие данные подвергаются дальнейшим манипуляциям: анализу и так далее.

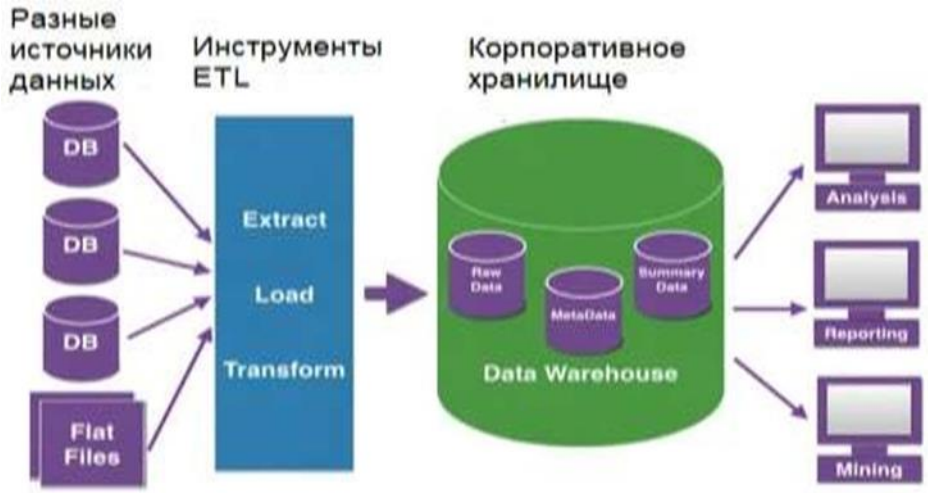

Рис.1. Данные извлекают, очищают и обрабатывают, помещают в корпоративное хранилище данных, а потом забирают для анализа.

Технологии BigData позволяют обработать большой объем неструктурированных данных, систематизировать их, проанализировать и выявить закономерности там, где человеческий мозг никогда бы их не заметил. Это открывает совершенно новые возможности по использованию данных. Само понятие BigData означает не просто большие данных. Это огромные хранимые и обрабатываемые массивы из сотен гигабайт, и даже петабайт данных. Данных, которые можно обработать и извлечь из них некоторое количество полезной информации. Говоря коротко, можно определить BigData как совокупность технологий обработки информации для получения информации.

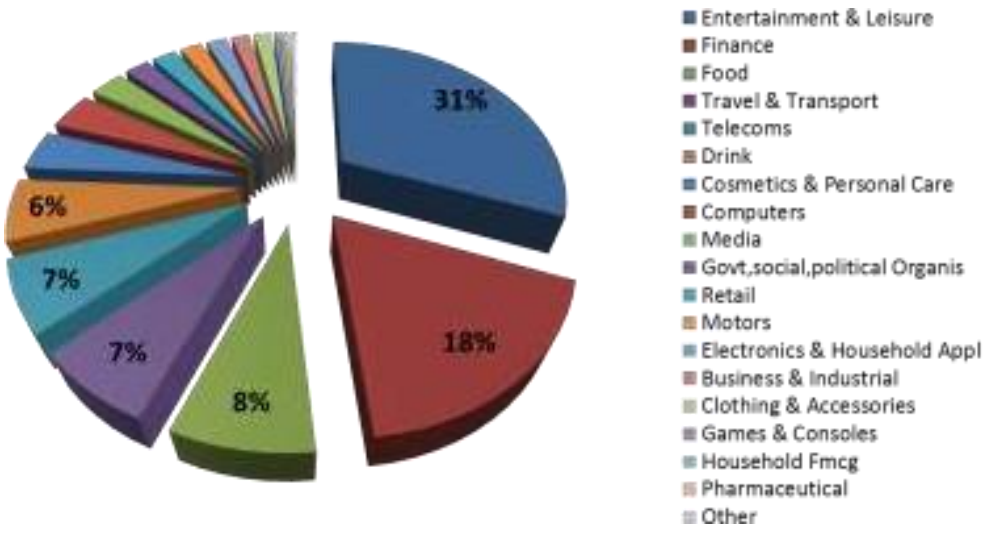

Рис. 2. Сфера применения Big Data

Технологии Big Data. Технологии, используемые для сбора и обработки BigData, можно разделить на 3 группы:

– Программное обеспечение;

– Оборудование;

– Сервисные услуги.

К наиболее распространенным подходам обработки данных (ПО) относятся:

SQL – язык структурированных запросов, позволяющий работать с базами данных. С помощью SQL можно создавать и модифицировать данные, а управлением массива данных занимается соответствующая система управления базами данных.

NoSQL – термин расшифровывается как Not Only SQL (не только SQL). Включает в себя ряд подходов, направленных на реализацию базы данных, имеющих отличия от моделей, используемых в традиционных, реляционных СУБД. Их удобно использовать при постоянно меняющейся структуре данных. Например, для сбора и хранения информации в социальных сетях.

MapReduce – модель распределения вычислений. Используется для параллельных вычислений над очень большими наборами данных (петабайты и более). Таким образом запрос представляет собой отдельную программу. Принцип работы заключается в последовательной обработке данных двумя методами Map и Reduce. Map выбирает предварительные данные, Reduce агрегирует их.

Hadoop – используется для реализации поисковых и контекстных механизмов высоконагруженных сайтов – Facebook, eBay, Amazon и др. Отличительной особенностью является то, что система защищена от выхода из строя любого из узлов кластера, так как каждый блок имеет, как минимум, одну копию данных на другом узле.

SAP HANA – высокопроизводительная NewSQL платформа для хранения и обработки данных. Обеспечивает высокую скорость обработки запросов. Еще одним отличительным признаком является то, что SAP HANA упрощает системный ландшафт, уменьшая затраты на поддержку аналитических систем.

Первыми технологии BigData стали применять те отрасли, деятельность которых завязана на обработке больших потоков информации ежедневно, – банки, мобильные операторы, торговые сети. В основном работа с данными в этих сферах направлена на формирование портрета клиента, чтобы предложить ему наиболее подходящие для него услуги.

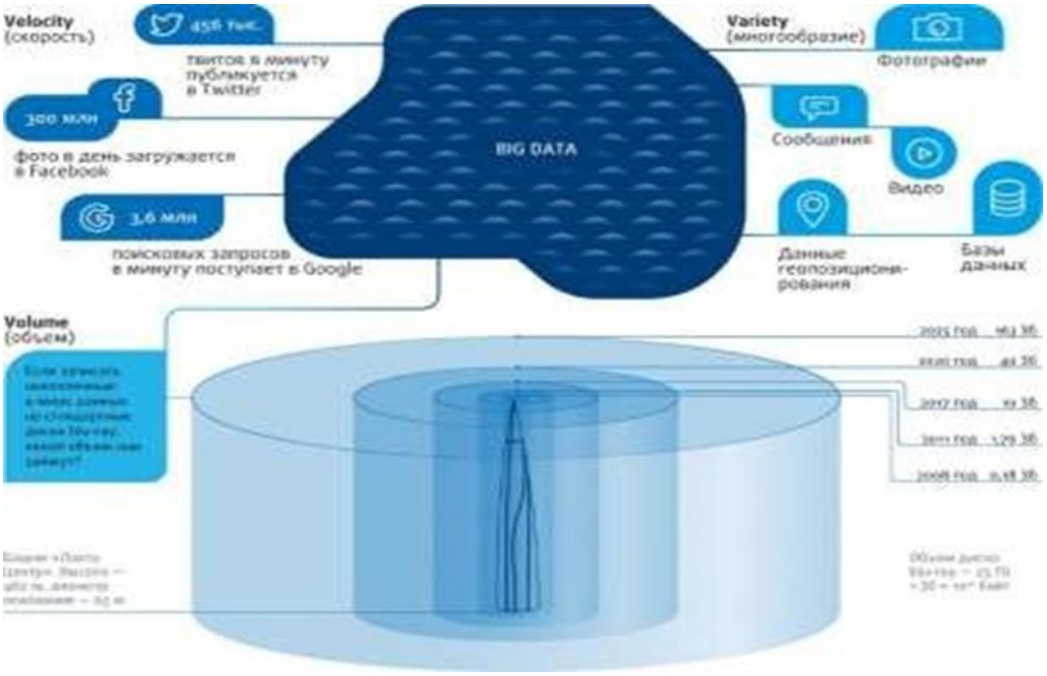

Рис. 3. Анализ контента с Big DATA в социальных сетях

Хранение и управление. Это как раз тот случай, когда приходится признать, что в BigData есть проблемы. Чем больше объем накопленных данных, тем требовательнее система хранения и управления этими данными. Вам придется покупать дорогостоящее оборудование или смирится с недостатками хранения данных в облаке. Вам понадобятся специалисты, способные предусмотреть возможные проблемы при анализе больших объемов данных, которые смогут организовать все нюансы таким образом, чтобы вы реально эффективно использовали данные.

Предвзятость. Предвзятость – еще одна из серьезных проблем в BigData. Довольно легко сделать конкретный вывод, если в вашем распоряжении результаты одного или двух исследований, но, если их становится значительно больше, появляется довольно большой простор для маневра, который позволяет изменить общий смысл результатов, изменив представление данных. Поэтому очень важно позаботиться о том, чтобы на результаты исследований не влияло мнение какой-либо из заинтересованных сторон.

Чем больше у вас данных, тем сложнее выделить именно то, что необходимо вам в текущий момент. Конечно, природа этой проблемы напрямую связана со спецификой BigData и вообще Data Mining, но ее не стоит упускать из виду.

«Каждый раз, когда вы пользуетесь поисковыми системами Google или Яндекс, вы работаете с большими данными».

Результаты и обсуждение. *Принципы работы с большими данными.* Исходя из определения Big Data, можно сформулировать основные принципы работы с такими данными:

Горизонтальная масштабируемость. Поскольку данных может быть сколь угодно

много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой. В 2 раза вырос объём данных – в 2 раза увеличили количество железа в кластере и всё продолжило работать.

Отказоустойчивость. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Hadoop-кластер Yahoo имеет более 42000 машин (по этой ссылке можно посмотреть размеры кластера в разных организациях). Это означает, что часть этих машин будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность таких сбоев и переживать их без каких-либо значимых последствий.

Локальность данных. В больших распределённых системах данные распределены по большому количеству машин. Если данные физически находятся на одном сервере, а обрабатываются на другом – расходы на передачу данных могут превысить расходы на саму обработку. Поэтому одним из важнейших принципов проектирования BigData-решений является принцип локальности данных – по возможности, обрабатываем данные на той же машине, на которой их храним.

Все современные средства работы с большими данными так или иначе следуют этим трём принципам. Для того, чтобы им следовать – необходимо придумывать какие-то методы, способы и парадигмы разработки средств разработки данных.

Например, пока еще не раскрыт весь потенциал больших данных в медицине. Алгоритмы машинного обучения уже активно применяются в диагностике онкологических заболеваний, но этот подход не используется в других областях, например, в лечении гриппа и персонализированных советов по диете.

Было бы интересно посмотреть на связку больших данных и дополненной реальности. Городские и музейные гиды, инструкции ко всему, что попадает в объектив вашей мобильной камеры, советы по первой помощи – сейчас просто не хватает фантазии, чтобы представить эффект синергии двух этих технологий в будущем. В будущем наш институт как полноценный партнер проекта ELBA (Создание учебных и исследовательских центров и разработка курсов по интеллектуальному анализу больших данных в Центральной Азии) в рамках программы Erasmus+, то мы будем повышать наши знания и применять эти технологии разные сферы нашей Республики.

Заключение. После прочтения этой статьи может показаться, что в анализе данных больше проблем, чем пользы, но не стоит забывать о том, что при умелом использовании это мощный и действенный инструмент, который способен помочь принимать эффективные решения. В частности, для грамотной работы с большими данными необходимо хорошо понимать специфику конкретного рынка и бизнеса, поэтому многие аналитики советуют создавать специалистов по анализу данных внутри компании.

Big Data открывает перед нами новые горизонты в планировании производства, образовании, здравоохранении и других отраслях. Если их развитие будет продолжаться, то технологии Big Data могут поднять информацию, как фактор производства, на совершенно новый качественный уровень. Информация станет не

только равноценна труду и капиталу, но и, возможно, станет важнейшим ресурсом современной экономики.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Виктор Майер-Шенбергер, Кеннет Кукьер “Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим” Москва; 2014.
2. Андреас Вайгенд BIG DATA. Вся технология в одной книге «Эксмо» 2017
3. Абдирахимов И. Э. (2021). Деэмульгирование нефтеводяных эмульсий. Universum: технические науки, (4-3 (85)), 72-75.
4. Билл Фрэнкс Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики, Москва; 2016
5. Халилов Ф. В. Облачные информационные технологии //инновации в информационных технологиях, машиностроении и автотранспорте. – 2017. – С. 179-181.
6. Абдирахимов, И. Э., Халимов, А. А., & Турсунов, Р. И. (2020). Подготовка качественного природного газа перед транспортировкой потребителю. Международный академический вестник, (2), 100-103.
7. Абдирахимов, И.Э., & Алиев, Ж. Ш. (2020). Технология бурения многоствольных скважин. Международный академический вестник, (2), 97-100.
8. Илхом Эшбоевич Абдирахимов, Шомансухрон Кароматходжа оглы Турасуннат, Азиз Тешабоевич Курбанов. Тепловые насосы для подогрева сетевой воды (2020). Science Time, 55-58.
9. Масьуд Убайдулла ўғли Каримов, Илхом Эшбоевич Абдирахмов (2022). Получение импортозамещающих диэмулгаторов на основе местного сырья. SCIENTIFIC PROGRESS.(2) 221-227.
10. Джураева, Г. Х., Абдирахимов, И.Э., & Шоназаров, Э.Б. (2021). Получение глауберовой соли и сульфата натрия из природного сырья. Universum: технические науки, (2-3 (83).
11. Виктор Майер-Шенбергер, Кеннет Кукьер. Большие данные. Революция, которая изменит то, как мы живём, работаем и мыслим Big Data. A Revolution That Will Transform How We Live, Work, and Think / пер. с англ. Инны Гайдюк. - М.: Манн, Иванов, Фербер, 2014. 240 с.