А.В. Свирновский,
магистрант 1 курса,
e-mail: ***vladimirovich1998@gmail.com,***
науч. рук.: ***В.Ф. Алексеев,***
к.т.н., доц.,
БГУИР,
г. Минск, Беларусь

МЕТОДЫ АНАЛИЗА И ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Аннотация: в данной статье рассматриваются основные методы анализа и обработки больших данных.

Ключевые слова: хранилище данных, большие данные, анализ данных, обработка данных

Повсеместное распространение технологий и доступа к Интернету приводят к бесконечному росту объема информации. 90% информации в этом мире было сгенерировано за два последних года. Сейчас, если огрублять, в день мы производим порядка 2,5 квинтильонов байт новой информации [1].

Рост количества информации привел к необходимости создания подходов к хранению и обработки такого огромного количества данных, а также к появлению термина «большие данные».

Компаниям необходимо было структурированно хранить, обрабатывать и анализировать всю полученную из внешних и внутренних источников информацию. Анализ больших данных проводят для того, чтобы получить новую, ранее неизвестную информацию.

Так начали появляться технологии BigData, позволяющие не только хранить и обрабатывать данные, но и снизить затраты на их хранение и обработку.

Примеры источников больших данных:

1. Логи поведения пользователей в интернете.

2. Джипиэс-сигналы от автомобилей для транспортной компании.

3. Данные, снимаемые с датчиков в большом адронном

коллайдере.

4. Оцифрованные книги в Национальной библиотеке РБ.

5. Информация о транзакциях всех клиентов банка.

6. Информация о всех покупках в крупной ритейл сети и т.д. [2].

Исходя из определения BigData, можно сформулировать основные принципы работы с такими данными:

1. 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 увеличится количество компонентов в кластере и всё продолжит работать.

2. Отказоустойчивость. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Хадуп-кластер Йаху имеет более 42000 машин. 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.

3. 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 обработка данных происходит на той же машине, на которой они хранятся [3].

Все современные средства работы с большими данными так или иначе следуют вышеупомянутым трём принципам. Для того чтобы им следовать необходимо определить методы, способы и парадигмы разработки средств разработки данных:

1. Методы класса или глубинный анализ (Дата Майнинг). Данные методы достаточно многочисленны, но их объединяет одно: используемый математический инструментарий в совокупности с достижениями из сферы информационных технологий.

2. Краудсорсинг. Данная методика позволяет получать данные одновременно из нескольких источников, причем количество последних практически не ограничено.

3. A/B-тестирование. Из всего объема данных выбирается контрольная совокупность элементов, которую поочередно сравнивают с другими подобными совокупностями, где был изменен один из элементов. Проведение подобных тестов помогает определить, колебания какого из параметров оказывают наибольшее влияние на контрольную совокупность. Благодаря объемам BigData можно проводить огромное число итераций, с каждой из них приближаясь к максимально достоверному результату.

4. Прогнозная аналитика. Специалисты в данной области стараются заранее предугадать и распланировать то, как будет вести себя подконтрольный объект, чтобы принять наиболее выгодное в этой ситуации решение.

5. Машинное обучение (искусственный интеллект). Основывается на эмпирическом анализе информации и последующем построении алгоритмов самообучения систем.

6. Сетевой анализ. Наиболее распространенный метод для исследования социальных сетей – после получения статистических данных анализируются созданные в сетке узлы, то есть взаимодействия между отдельными пользователями и их сообществами [2].

Одним из пользующихся спросом программным обеспечением обработки больших данных является платформа Майкрософт Ажур Плэтформ.

Майкрософт Ажур Плэтформ – это постоянно расширяющийся набор облачных сервисов, которые помогают любой организации решать бизнес-задачи. Это свобода создания, управления и развертывания приложений в огромной глобальной сети с использованием любимых инструментов и сред.

Преимущества Майкрософт Ажур Плэтформ:

1. Продуктивность. Возможность сокращения маркетинговых циклов, предоставляя функции быстрее с более чем 100 сквозными услугами.

2. Гибридность. Возможность разрабатывать и

разворачивать в любом месте, используя единственное на рынке гибридное облако.

3. Разумность. Возможность создавать интеллектуальные приложения, используя мощные службы данных и искусственного интеллекта.

4. Доверие. Возможность присоединения к стартапам, правительствам и 95 процентам компаний из списка Форчен 500, которые сегодня работают в облаке Майкрософт.

Ажур Плэтформ обеспечивает несколько моделей развертывания и сервисов для обработки больших данных. Эти варианты сервисов для обработки больших данных позволяют начать с уровня затрат и возможностей, соответствующих необходимому случаю использования, а затем предоставляют гибкие возможности изменения выбора по мере изменения ваших требований.

Литература и примечания:

[1] Технологии и средства связи [Электронный ресурс]. Режим доступа: http://tssonline.ru/articles2/fix-corp/rost-obema-informatsii--realii-tsifrovoy-vselennoy – Дата доступа: 01.02.2020.

[2] Черняк Л. Большие Данные – новая теория и практика // Открытые системы. СУБД. – 2011. – №10. – С. 18-25.

[3] Большие данные [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Большие_данные – Дата доступа: 02.02.2020.