УДК 004.67

ОБРАБОТКА СВЕРХБОЛЬШИХ МАССИВОВ ДАННЫХ
♦♦♦♦♦
BIG DATA ARRAYS PROCESSING

Наталочка Анастасия Денисовна
студент,
Кубанский государственный технологический университет
asnatalochka@gmail.com

Отришко Александр Александрович
студент,
Кубанский государственный технологический университет
otrishko.san@mail.ru

Янаева Марина Викторовна
Кандидат технических наук, доцент
кафедры информационных систем и программирования,
Кубанский государственный технологический университет
yanaevam@mail.ru

Аннотация. Данная статья посвящена обзору технологии «большие данные» (BigData) а также ее отличительных черт. Обоснованы необходимость использования и перспективность применения технологий BigData. Проведен анализ существующих программно-аппаратных средств, использующихся для анализа и обработки больших данных, таких как MapReduce, Hadoop и HBase, выделены их преимущества и особенности.

Ключевые слова: большие данные, информация, горизонтальная масштабируемость, статистический анализ, отказоустойчивость, MapReduce, Hadoop, HBase.

Natalochka Anastasia Denisovna
Student,
Kuban State Technological University
asnatalochka@gmail.com

Otrishko Alexander Alexandrovich
Student,
Kuban State Technological University
otrishko.san@mail.ru

Yanaeva Marina Viktorovna
Candidate of Technical Sciences,
Associate Professor of the Department
of Information Systems and Programming,
Kuban State Technological University
yanaevam@mail.ru

Annotation. This article is devoted to an overview of the Big Data technology as well as its distinctive features. The necessity of using and the prospects of using Big Data technologies are grounded. The analysis of existing software and hardware used for the analysis and processing of Big Data, such as MapReduce, Hadoop and HBase, was carried out, their advantages and features were highlighted.

Keywords: big data, information, horizontal scalability, statistical analysis, fault tolerance, MapReduce, Hadoop, HBase.

Введение. Объем данных, генерируемый и собираемый современными научно-исследовательским центрами, финансовыми институтами, социальными сетями, уже привычно измеряется петабайтами.

Таким образом, в современном мире возникла проблема больших данных или BigData. Мировые лидеры в сфере ИТ и бизнеса заняты поиском оптимального решения для управления огромным количеством постоянно прибывающей информации и ее анализа.

Тема больших данных интересна как с практической, так и с теоретической точек зрения. Сами технологии находятся в состоянии непрерывного развития, что позволяет как в режиме реального времени наблюдать за процессом их внедрения и совершенствования, так и непосредственно участвовать в создании новых технологий обработки больших массивов данных.

Что такое большие данные. Из названия можно предположить, что термин «большие данные» относится просто к управлению и анализу больших объемов данных.

Тем не менее «большие данные» предполагают нечто большее, чем просто анализ огромных объемов информации. Проблема не в том, что организации создают огромные объемы данных, а в том, что большая их часть представлена в формате, плохо соответствующем традиционному структурированному формату БД.

BigData – это термин, описывающий наборы данных колоссальных объёмов и размеров (они могут быть 100 Гб, 1 ТБ, а могут быть настолько большими что числами не измерить), очень быстро растущие с течением времени, а также инструменты для обслуживания и работы с ними. Это отличный способ собрать и обработать огромное количество информации, чтобы решать довольно трудные прикладные задачи, которые решить без использования BigData сложно, а то и невозможно.

По сути, большие данные – довольно условное и относительное понятие. Самое распространенное его определение – это набор информации, по объему превосходящей жесткий диск одного персонального устройства и не поддающейся обработке классическими инструментами, применяемыми для меньших объемов.

Характеристики BigData. Есть характеристики, которые позволяют отнести информацию и данные именно к BigData. То есть не все данные могут быть пригодны для аналитики. В этих характеристиках как раз и заложено ключевое понятие биг дата.

Описывая специфику больших данных, первым делом упоминают 3V: «volume, variety and velocity» или объем, разнообразие и скорость.

Объем подразумевает не только большое количество хранимой информации, но и ее дополнение, рост, изменение с течением времени.

Разнообразие типов и источников информации всегда было большой проблемой, когда появлялась необходимость свести их в один массив данных. Сегодня это разнообразие только увеличивается.

Скорость оценивается как при создании информации, так и при ее обработке.

Однако периодически к VVV добавляют и четвертую V (veracity – достоверность/правдоподобность данных) и даже пятую V (в некоторых вариантах это – viability – жизнеспособность, в других же это – value – ценность).

Принципы работы с большими данными. Работать с BigData можно несколькими способами: Горизонтальная масштабируемость, отказоустойчивость, локальные данные. Рассмотрим отдельно каждый из способов.

1. Горизонтальная масштабируемость. Так как данные могут быть неограниченного количества и неограниченного размера, то и система, которая работает с обработкой больших данных, должна быть расширяемой. Если увеличивается объем информации, который необходимо обрабатывать, то и должно увеличиваться количество и мощность системы, которая их будет обрабатывать и работать с ними. Кластер должен работать, не переставая, при любом количестве информации необходимой для работы.

2. Отказоустойчивость. Принцип горизонтальной масштабируемости говорит о том, что мощности системы неограниченны и все время увеличивается. То есть в обычной системе если хотя бы одна машина выйдет из строя, то вся система обрушится и перестанет работать до ее восстановления. Система, работающая с большими данными должна быть спасена от таких ситуаций и предусмотреть что неопределенное количество машин может в любой момент выйти из строя.

3. Локальность данных. Количество данных и их размер огромны. Если хранить все данные на одной машине и передавать для обработки на другую, то на это уйдет огромное количество ресурсов и времени. Поэтому на машинах должно храниться столько информации и такого типа, которые она обрабатывает. То есть где хранится, там и обрабатывается.

Техники и методы анализа, применимые к BigData

- слияние и интеграция данных (Data fusion and data integration);
- интеллектуальный анализ данных (Datamining);
- машинное обучение (machinelearning);
- генетические алгоритмы (Genetic algorithms);
- нейронные сети (Neural networks);
- обработка потоков (Stream processing) и др.

Технологии работы с большими данными. Большинство аналитиков относит к технологиям обработки и анализа больших данных следующие средства:

- MapReduce;
- Hadoop;
- NoSQL.

MapReduce – это модель распределенной обработки данных, обрабатывающая большие объемы данных на кластерах компьютеров. Принцип работы MapReduce можно представить по приведенному ниже рисунку 1.

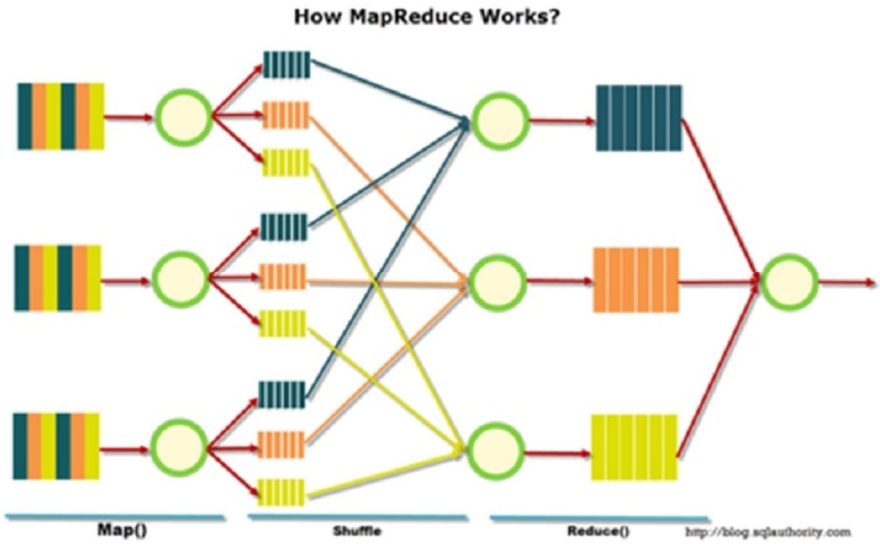

Рисунок 1 – Схема работы MapReduce

Этот способ предполагает, что данные представлены в виде некоторых записей. Их обработка проходит в 3 определенных этапа.

1. *Стадия Map*. На данной стадии данные получаются и предоставляются при помощи функции map(), которую определяет и создает пользователь. Работа этой стадии происходит по принципу предобработки и фильтрации, получаемых данных. Именно так работает функция map в функциональных языках программирования – то есть функция, которую задает пользователь, выполняется каждый раз при получении данных.

2. *Стадия Shuffle*. Эту стадию пользователи обычно не замечают, так как в этот момент вывод функции map передается в разные, так называемые, «корзины» с информацией. Каждая эта «корзина» соответствует одному ключу, выводимому в стадии map. В дальнейшем эти «корзины» передаются на вход для выполнения стадии reduce;

3. *Стадия Reduce*. Каждая «корзина» со значениями, сформированная на стадии выполнения shuffle, попадает в функцию reduce(). Функция reduce, так же, как и map, задается пользователем и вычисляется результат отдельной «корзины». Большая часть значений, которые возвращает функция reduce() после выполнения, является финальной стадией работы MapReduce.

Hadoop. Изначально, он был инструментом для хранения данных и запуска MapReduce задач, но сейчас Hadoop представляет собой большой стек технологий, связанных с обработкой больших данные, не только при помощи MapReduce.

Основные компоненты Hadoop: Hadoop Distributed File System (сокращенно HDFS), представляющую собой файловую систему, которая позволяет практически неограниченное количество данных с неограниченным объемом. HadoopYARN, представляющий собой фреймворк управляющий ресурсами кластеров и распределяя задачи по ним, в том числе выполнение MapReduce. Hadoopcommon. Так же есть не основные компоненты: Hive – помогающий выполнять SQL запросы над большими данными, Pig – язык программирования для выполнения анализа данных на высоком уровне, Hbase – база данных, реализующая парадигму BigTable, Cassandra – распределенная key-value база данных, ZooKeeper – сервис для хранения конфигурации и синхронизации изменения этой конфигурации, Mahout – библиотека и движок машинного обучения на больших данных.

Преимущества решения на базе Hadoop :

– снижение времени на обработку данных;
– снижение стоимости оборудования;
– повышение отказоустойчивости. технология позволяет построить отказоустойчивое решение;
– линейная масштабируемость;
– работа с неструктурированными данными.

NoSQL и HBase. Традиционные СУБД ориентируются на требования ACID к транзакционной системе: атомарность, согласованность, изолированность, надёжность, тогда как в NoSQL вместо ACID может рассматриваться набор свойств BASE:

– базовая доступность (англ. *basicavailability*) – каждый запрос гарантированно завершается (успешно или безуспешно);

– гибкое состояние (англ. *softstate*) – состояние системы может изменяться со временем, даже без ввода новых данных, для достижения согласования данных;

– согласованность в конечном счёте (англ. *eventualconsistency*) – данные могут быть некоторое время рассогласованы, но приходят к согласованию через некоторое время.

Решения NoSQL отличаются не только проектированием с учётом масштабирования. Другими характерными чертами NoSQL-решений являются:

– применение различных типов хранилищ;

– возможность разработки базы данных без задания схемы;

– линейная масштабируемость (добавление процессоров увеличивает производительность);

– инновационность: «не только SQL» открывает много возможностей для хранения и обработки данных.

Сейчас мы рассмотрим HBase и его возможности. HBase – СУБД класса NoSQL с открытым исходным кодом, проект экосистемы Hadoop.

HBase представляет собой попытку объединить пакетную обработку и удобство обновления файлов, также имея произвольный доступ к ним.

Она является распределенной, колоночно-ориентированной и мультиверсионной базой типа ключ-значение. Данные в нем организованы в таблицах, они проиндексированы по первичному ключу, называемые RowKey.

Заключение. Анализ и обработка больших данных – непростая и комплексная задача, требующая для решения особых инструментов и больших вычислительных возможностей. В их основе лежат математические алгоритмы, теория вероятностей и многие другие инструменты, которые при применении к большим данным могут принести больше плоды тем, кто не обошёл вниманием это относительно новое явление в информационном интернет-пространстве. Учитывая стремительный рост объёма больших данных, можно с достаточной уверенностью предполагать, что направления науки, связанные с их анализом, не потеряют актуальность в обозримом будущем.

Литература

1. Медетов А.А. Термин BigData и способы его применения // Молодой ученый. – 2016. – № 11. – С. 207–210.
2. Иванов П.Д., Вампилов В.Ж. Технологии BigData и их применение на современном промышленном предприятии // Инженерный журнал: Наука и инновации. – 2014. – Т. 8.
3. BigData от А до Я [Электронный ресурс]. – URL : https://habr.com/ru/post/267361/ (дата обращения 23.07.2022)
4. Назаренко Ю.Л. Обзор технологии «большие данные» (BigData) и программно-аппаратных средств, применяемых для их анализа и обработки [Электронный ресурс]. – URL : https://cyberleninka.ru/article/n/obzor-tehnologii-bolshie-dannye-big-data-i-programmno-apparatnyh-sredstv-primenyaemyh-dlya-ih-analiza-i-obrabotki (дата обращения 10.07.2022).

References

1. Medetov A.A. The term Big Data and methods of its application // Young scientist. – 2016. – № 11. – P. 207–210.
2. Ivanov P.D., Vampilov V.Zh. Big Data technologies and their application in a modern industrial enterprise // Engineering Journal: Science and Innovations. – 2014. – Vol. 8.
3. Big Data from A to Z [Online]. – URL : https://habr.com/ru/post/267361/ (accessed 23.07.2022)
4. Nazarenko Yu.L. Overview of Big Data technology and software and hardware used for their analysis and processing [Online]. – URL : https://cyberleninka.ru/article/n/obzor-tehnologii-bolshie-dannye-big-data-i-programmno-apparatnyh-sredstv-primenyaemyh-dlya-ih-analiza-i-obrabotki (accessed 07.10.2022).