

СОВРЕМЕННЫЕ МЕТОДЫ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ В КРУПНОМАСШТАБНЫХ СИСТЕМАХ

Качалов Д.Л., Мишустин А.В., Фархадов М.П.

Институт проблем управления РАН имени В.А. Трапезникова

Аннотация: в данной статье рассмотрены современные методы и модели обработки больших данных в крупномасштабных системах, в том числе и экономических системах. Приведены алгоритмы обработки данных, на основе которых можно анализировать и прогнозировать конъюнктуру рынка.

Ключевые слова: большие данные, большой андронный коллайдер, массивы данных, обработка данных.

Под большими данными понимается широкое разнообразие массивов данных, которые не могут быть надлежащим образом обработаны традиционными приложениями из-за своего огромного объема или сложного состава. Сложность анализа больших данных заключается в специфике их сбора, курирования, разделения, хранения, передачи, визуализации и сохранении конфиденциальности информации. Под анализом больших данных часто понимается применение прогнозной аналитики или других передовых методов с целью извлечения из множества данных определенной полезной информации. Точность при анализе больших данных помогает принимать более рациональные решения. В свою очередь, принятие наилучших решений позволяет увеличить производственную эффективность, сократить расходы и снизить риски.

Анализ больших данных может применяться в таких областях, как отслеживание конъюнктуры рынка, предотвращение распространения эпидемий и борьба с преступностью. Ученые, руководители крупнейших компаний, работники сферы масс-медиа и рекламы, а также правительственные органы часто сталкиваются с трудностями при анализе массивов данных огромных объемов в таких областях, как поиск в сети интернет, информационные технологии в сфере бизнеса и финансов и т.п. Работа ученых, особенно метеорологов, медиков, изучающих геномы, исследователей, работающих в области изучения средств коммуникации,

физиков, создающих сложные симуляторы, а также биологов и экологов часто ограничивается возможностями обработки огромных массивов данных.

Объемы массивов данных постоянно увеличиваются, так как расширяются возможности сбора информации при помощи недорогостоящих мобильных устройств, средств цифровой аэрофотосъемки, камер, микрофонов, считывателей радиочастотных меток и беспроводных сенсорных сетей. Общий объем информации, который может сохраняться всеми мировыми техническими средствами, удваивался примерно каждые 40 месяцев на протяжении 80-х годов XX века, а с 2012 года и по настоящее время ежедневно генерируется 2,5 экзабайтов ($2,5 \times 10^{18}$) информации.

1. Анализ и развитие технологий обработки больших данных в научных исследованиях

Рассмотрим, как используются технологии обработки больших данных в научных исследованиях.

При помощи Большого адронного коллайдера со 150 миллионов датчиков снимались данные 40 миллионов раз в секунду. Это около 600 миллионов столкновений в секунду. После фильтрации и удаления более 99,99995% из них осталось 100 столкновений в секунду, которые были изучены исследователями.

В результате, после изучения менее чем 0,001% данных, полученных от датчиков, объем данных со всех четырех экспериментов с использованием Большого адронного коллайдера составил 25 петабайт в год до репликации (по данным на 2012 год) или почти 200 петабайт после репликации.

Если бы фиксировались все данные, полученные от датчиков, установленных в Большом адронном коллайдере, то такой объем данных было бы чрезвычайно сложно обработать. Объем потока данных превысил 150 миллионов петабайт в год или 500 экзабайт в день до репликации. Это эквивалентно 500 квинтильонам (5×10^{20}) байт в день, что почти в 200 раз превышает объем данных, получаемых ото всех других ресурсов в мире.

1.1 Анализ методов и подходов к технологиям обработки больших данных

Под анализом больших данных понимается как анализ массивов данных в рамках возможностей персонального компьютера, так и в рамках возможностей систем управления реляционными базами данных, при этом как в первом, так и во втором случае при формировании и статистики, и визуализации возникают определенные трудности, которые заключаются в необходимости обеспечения скоординированной работы компьютерных программ на десятках, сотнях или даже тысячах серверов.

Анализ больших данных может быть охарактеризован по следующим параметрам:

1. Объем, т.е. количество генерируемых данных. От этого показателя зависит, может ли определенный массив данных считаться большими данными или нет. Данных хранятся SQL-серверах в облачной среде.
2. Многообразие, т.е. категория, к которой принадлежат большие данные. Знание такой принадлежности позволяет аналитикам наиболее эффективно работать с информацией.
3. Скорость, т.е. скорость генерирования или обработки данных с целью осуществления поставленных целей.
4. Изменчивость, т.е. нестабильность данных во времени.
5. Достоверность, т.е. качество собранных данных, от которого зависит точность анализа.
6. Сложность, т.е. трудоемкость процесса корреляции и построения взаимосвязей между данными.

Рассмотрим основные методы анализа больших данных. Методы анализа больших данных, применяемых в современных технологиях, можно отобразить с помощью следующей диаграммы (рисунок 1).

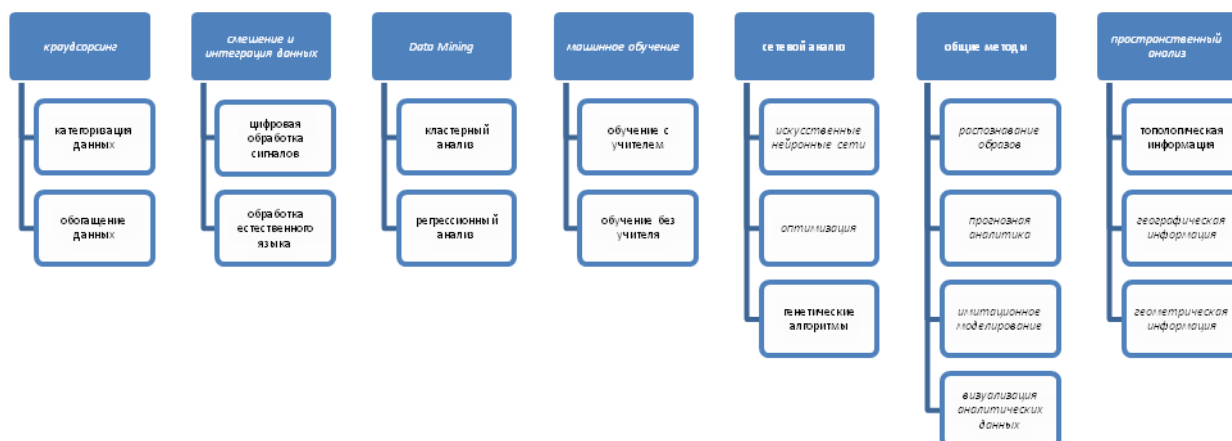


Рисунок 1 – Методы анализа больших данных

1.2 Архитектура технологий обработки больших данных

В 2004 году Google опубликовал статью, посвященную процессу MapReduce. Каркас MapReduce использует модель параллельной обработки очень большого объема данных. В рамках MapReduce запросы разделяются и распределяются по параллельным узлам, при этом они обрабатываются параллельно (этап распределения (Map)). Затем осуществляется сбор и доставка результатов (этап предварительного преобразования (Reduce)). Этот каркас был весьма эффективен, поэтому другие компании также захотели скопировать алгоритм. Каркас MapReduce был использован в проекте Apache с открытым исходным кодом, носящем название Hadoop.

Последние исследования показали, что многоуровневая архитектура может успешно использоваться для анализа больших данных. Архитектура распределенной параллельной обработки распределяет данные по многочисленным узлам, которые, производя, соответственно, параллельную обработку, генерируют результат гораздо быстрее, чем другие системы. В рамках такой архитектуры данные передаются в параллельные СУБД с использованием каркасов MapReduce и Hadoop. При этом задействуются фронтальные серверы приложений.

2. Решение практических задач с помощью технологии обработки больших данных

Обработка больших данных сейчас предполагает, как правило, внедрение специальных программных комплексов, таких как, Hadoop, позволяющих производить обработку больших объёмов данных на основании концепции Map-Reduce.

Hadoop на данный момент является «де-факто» стандартом обработки больших данных. Hadoop представляет собой фреймворк, на основе которого разрабатываются приложения для анализа и визуализации больших данных.

Хранение данных в данном фреймворке осуществляется с помощью специальной распределённой файловой системы HDFS (Hadoop Distributed File System), которая лежит в основе Hadoop и позволяет хранить и предоставлять доступ к данным сразу на нескольких узлах кластера. Таким образом, если один или несколько узлов кластера выходят из строя, то риск потери информации сводится к минимуму и кластер продолжает работу в штатном режиме.

Процесс обработки данных в Hadoop-приложении

Для обработки больших данных используется алгоритм Map-Reduce, при этом все стадии Map должны завершить свою работу до начала Reduce. Также входные данные требуют предварительной обработки. Таким образом, получается общий алгоритм работы, изображенный на рисунке 2.

Одной из самых актуальных задач современных информационных технологий является задача быстрой обработки больших объёмов данных. Эффективное решение данной задачи, позволяет быстрее принимать решения на основе данных, полученных в прошлом. В работе были проанализированы методы и подходы к технологии обработке больших данных.

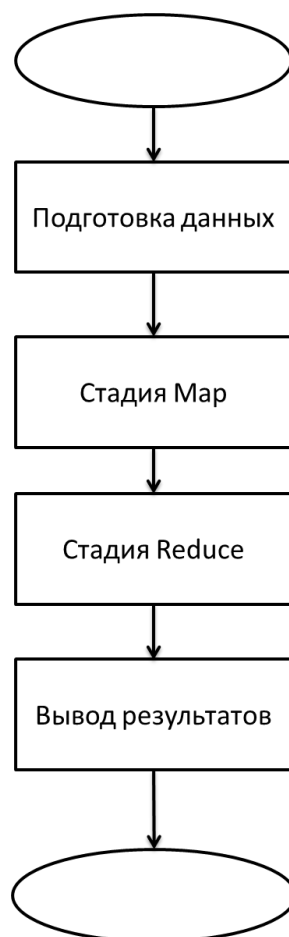


Рисунок 2 – Алгоритм анализа данных

Список использованных источников

1. *Franks, B. Taming the Big Data Tidal Wave Finding Opportunities in Huge Data Streams with Advanced Analytics/ Bill Franks, 2012. – 45 с.*
2. *Gantz, J. The digital universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East - United States / J. Gantz, D. Rainsel // IDC Country brief, 2013.*
3. Hadoop and Big Data [Электронный ресурс] – Режим доступа : <http://www.cloudera.com/content/cloudera/en/about/hadoop-and-big-data.html>.

MODERN METHODS OF BIG DATA ANALYSIS IN LARGE-SCALE SYSTEMS

Kachalov D.L., Mishustin A.V., Farhadov M.P.

Russia, V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences

Abstract: In this paper, modern methods of big data analysis in large-scale systems was described. Algorithms of data analysis was described, which based at financial analysis methods.

Keywords: big data, data analysis.