Перспективы Науки и Образования. 2016. 3 (21)

Международный электронный научный журнал
ISSN 2307-2334 (Онлайн)

Адрес статьи: pnojournal.wordpress.com/archive16/16-03/
Дата публикации: 1.07.2016
№ 3 (21). С. 7-11.
УДК 004.041

Е. Е. Чехарин

Большие данные: большие проблемы

Статья описывает особенности проблемы «больших данных». Анализируются причины периодического возникновения проблемы. Показано, что проблема существовала задолго до ее отражения в средствах массой информации. Описаны основные качественные характеристики больших данных: большой объем, сложность, временные ограничения. Описаны числовые характеристики больших данных. Описаны методики и методы работы с большими данными.

Ключевые слова: большие данные, информационные технологии, информационные объемы данных, методы обработки, сложность систем данных, время обработки данных, анализ данных, моделирование

Perspectives of Science & Education. 2016. 3 (21)

International Scientific Electronic Journal
ISSN 2307-2334 (Online)

Available: psejournal.wordpress.com/archive16/16-03/
Accepted: 1 June 2016
Published: 1 July 2016
No. 3 (21). pp. 7-11.

E. E. Chekharin

Big data: big problems

This article describes the contents of the problem of "big data." The article analyzes the causes of the problem batch. The article shows that the problem existed long before its reflection in the media. The article describes the main qualitative characteristics of Big Data: high volume, complexity and time constraints. This article describes the characteristics of large numerical data. This article describes the techniques and methods of working with big data.

Keywords: big data, IT, information data volumes processing methods, the complexity of these systems, while data processing, data analysis, modeling

Введение

Проблема «больших данных» (BigData) [1, 2] начала обсуждаться с 2008 г. Эту проблему чаще всего связывают с необходимостью анализа неструктурированных данных больших объёмов. Для характеристики «больших данных» ввели упрощенные критерии, которые назвали «три V»: объём (volume), скорость (velocity), многообразие (variety). Это означает, что большие данные появляются при наличие большого объема (volume), который является проблемным для средств обработки. Это означает, что большие данные появляются при требовании быстрой обработки или высокой скорости обработки (velocity) , которую не могут обеспечить средства обработки. Это означает, что большие данные появляются при высокой сложности [3] или упрощенно разнообразие (variety) , которую не могут исследовать методы анализа и обработать средства обработки. Появление термина «большие данные» связывают с 2008 годом[4]. Введение термина «большие данные» связывают с Клиффордом Линчем – редактором журнала Nature [4], подготовившему серию работ на эту тему. Это обозначает признание проблемы в некомпьютерных сферах.

Проблему больших данных выявили специалисты в области дистанционного зондирования Земли более 60 лет назад [5], спустя некоторое время ее отметили программисты. Затем ее зафиксировали аналитики 20-30 лет назад. И только в последние десять лет она открылась для бизнес-аналитиков и журналистов, что и привело к их повышенному вниманию к такому явлению и появлению термина «большие данные».

В процессе развития человеческого общества происходит наблюдение человека за объектами,

явлениями и процессами окружающего мира. Как результат наблюдения происходит получение информации в информационном поле [6, 7], накопление опыта и формирование описаний объектов, явлений и процессов. Первичное описание объектов окружающего мира состояло в формировании количественных и качественных свойств, характеристик, признаков и отношений между ними. Это описание представляет собой информационные коллекции. Вторичное описание состояло в формировании моделей и систем, формируемых на основе анализа первичных коллекций данных. Чем сложнее объект исследования, тем большего количества информации требует его описание и тем объемнее и сложнее информационные коллекции, составляющие такое описание.

Рост объемов собираемой информации и требование ее обработки и хранения делают актуальным исследования в области методов и алгоритмов анализа больших и сверхбольших коллекций данных. В работе [8] высказана гипотеза о том, что выявление закономерностей в больших массивах данных становится одним из инструментов исследования и одним из методов получения новых знаний в современных условиях. Если в прежнее время появление новых фактов легко фиксировалось и становилось предметом исследования, то в настоящее время проблемой становиться нахождение таких новых фактов и их формализация в больших массивах данных.

Один из признаков больших данных большой объем информационных коллекций характеризует как IT-компании, так и научную сферу [9], а также широкий спектр организаций в самых различных областях. Поэтому в современной науке возникло новое научное направление, связанное с анализом больших и сверхбольших наборов данных - BigData [2].

Характеристики больших данных

Проблемы больших данных, применяемых в разных прикладных областях требуют проведения исследований и разработок, направленных на создание масштабируемых аппаратных и программных решений проблем. Пока пределом возможностей современных программных приложений, ориентированных на обработку больших объемов данных, являются петабайтные наборы и гигабайтные потоки данных. Но в соответствии с тенденцией развития науки и общества ожидаются еще большие масштабы и объемы данных.

При создании приложений, работающих с большими данными, приходится сталкиваться со следующими характеристиками: большие объемы данных, интенсифицированные потоки данных, высокая структурная сложность, нелинейность моделей, требование существенного сокращения времени анализа или обработки данных, предел времени принятия решений при любом количестве данных [10], возрастание морфологической сложности моделей, возрастание структурной сложности моделей и систем, возрастание когнитивной сложности [11], рост слабоструктурированной исходной информации, относительный рост нечеткой информации, рост потребностей в параллельных вычислениях [12] и т.д.

Схематически проблемы работы с данными большого объема приведены в таблице 1.

Таблица 1

Сравнительные характеристики обычных и больших данных

Характеристика	Обычные данные	Большие данные
Формат	Однородный	Неоднородный
Объем	Мегабайты гигабайты	Петабайты
Распределенность данных	нет	есть
Тип задачи	Первого рода	Второго рода
Тип моделей решателей	Алгоритмические	Статистические
Тип моделирования	Имитационное моделирование	Стохастическое
Топологическая сложность	Приемлемая	Высокая
Вычислительные ресурсы	Обычные	Повышенной мощности

Программы, ориентированные на обработку больших объемов данных, имеют дело с файлами данных объемом от нескольких терабайт до петабайта. На практике эти данные поступают в разных форматах и часто распределены между несколькими источниками хранения информации. Обработка подобных наборов данных обычно происходит в режиме поэтапного аналитического конвейера, включающего стадии преобразования и интеграции данных.

Требования к объему и скорости вычислений обычно линейно возрастают при росте объема данных. Простейший подход основан на использовании распараллеливания большого объема данных. К основным исследовательским проблемам относятся управление такими данными, методы фильтрации и интеграции данных, эффективная поддержка запросов и распределенности данных.

Особо следует отметить множественность форматов данных, которая сама по себе создает проблемы даже при не очень большом объеме. Это мотивирует разработку специальных информационной конструкции [13] и моделей информационных взаимодействий [14], которые часто отображают свойства информационного пространства или свойства поля [7].

Для приложений, ориентированных на обработку больших объемов данных, характерны возрастающая вычислительная сложность. Требования к вычислениям нелинейно возрастают при росте объемов данных; для обеспечения правильного вида данных требуется применение сложных методов поиска и интеграции. Ключевыми исследовательскими проблемами являются разработка новых алгоритмов, генерация сигнатур данных и создание специализированных вычислительных платформ, включающих аппаратные ускорители.

К числу приложений, которым свойственны соответствующие характеристики, относятся следующие. A/B testing. Методика, в которой контрольная выборка поочередно сравнивается с другими. Тем самым удается выявить оптимальную комбинацию показателей для достижения, например, наилучшей ответной реакции потребителей на маркетинговое предложение. Большие данные позволяют провести огромное количество итераций и таким образом получить статистически достоверный результат.

Ad-hoc GRID - непосредственное формирование сотрудничающих гетерогенных вычислительных узлов в логическое сообщество без предварительно сконфигурированной фиксированной инфраструктуры и с минимальными административными требованиями.

Association rule learning. Набор методик для выявления взаимосвязей, т.е. ассоциативных правил, между переменными величинами в больших массивах данных. Используется в data mining.

BOINC-грид. Как правило, для обработки больших массивов данных используются суперкомпьютеры или вычислительные кластеры. Для достижения большей производительности вычислительные кластеры объединяются высокоскоростными каналами связи в специализированные ГРИД-системы. Однако с развитием сети Интернет появился и другой подход в построении ГРИД-систем, позволяющий объединить значительное число источников сравнительно небольших вычислительных ресурсов для решения задач обработки больших и сверхбольших объемов данных. В большинстве случаев такие системы построены на использовании свободных вычислительных ресурсов частных лиц и организаций, добровольно присоединяющихся к этим системам (volunteer computing). Однако существуют и примеры построения подобных частных (в масштабах организации или группы организаций) распределенных систем [15]. Наиболее эффективно использование таких распределенных систем для проведения серий независимых вычислительных экспериментов. Calculation acceleration - ускорение вычислений - изменение скорости вычислений в одной системе при сравнении со скоростью вычислений в другой системе.

Набор методик, которые позволяет предсказать поведение потребителей в определенном сегменте рынка (принятие решений о покупке, отток, объем потребления и проч.) - Classification. Используется в интеллектуальном анализе данных.

Пакет Global GRID (глобальные ГРИД) - устанавливается в Интернете, предоставляя отдельным пользователям или организациям мощность ГРИД независимо от того, где в мире эти пользователи находятся. Это также называют Интернет-компьютингом. Другой пакет Cluster analysis использует классический метод классификации объектов по кластерам за счет выявления априори не известных общих признаков. Используется в data mining.

Cluster - кластер - доступная по сети группа рабочих узлов (при необходимости вместе с головным узлом), размещённая на некотором сайте. Согласно определению в схеме GLUE, кластер это контейнер, который группирует вместе подкластеры или компьютерные узлы.

Cluster and multi-cluster GRIDs model – кластерная и мультикластерная модель ГРИД. Crowdsourcing. Методика сбора данных из большого количества источников с последующей кластеризацией по неоднорозным признакам.

Data GRID - проект, финансируемый Европейским Союзом. Цель проекта - создание следующего поколения вычислительной инфраструктуры обеспечения интенсивных вычислений и анализа общих крупномасштабных баз данных (от сотен терабайт до петабайт) для международных научных сообщений.

Data fusion and data integration.Набор методик, который позволяет анализировать комментарии пользователей социальных сетей и сопоставлять с результатами продаж в режиме реального времени.

Когнитивная кластеризация [16]. Универсальный метод кластеризации, основанный на включение когнитивной области человека в анализ на предварительной стадии. Ensemble learning. В этом методе задействуется множество предикативных моделей за счет чего повышается качество сделанных прогнозов. Genetic algorithms. В этой методике возможные решения представляют в виде `хромосом`, которые могут комбинироваться и мутировать. Как и в процессе естественной эволюции, выживает наиболее приспособленная особь.

GRID (грид, сеть) - географически распределенная информационная система (не путать с геоинформационной системой)- технология распределённых вычислений, в которой вычислительная система («суперкомпьютер») представлена в виде соединенных сетью вычислительных узлов, слабосвязанных, гомогенных или

гетерогенных компьютеров, работающих вместе для выполнения большого количества заданий. ГРИД-технология применяется для решения разного рода научных задач, требующих значительных вычислительных ресурсов.

GRID infrastructure - инфраструктура ГРИД - географически распределённая инфраструктура, объединяющая множество ресурсов разных типов (процессоры, долговременная и оперативная память, хранилища и базы данных, сети), доступ к которым пользователь может получить из любой точки, независимо от места их расположения.

Machine learning. Направление в информатике (исторически за ним закрепилось название `искусственный интеллект`), которое преследует цель создания алгоритмов самообучения на основе анализа эмпирических данных.

MIMD, Multiple Instruction Multiple Data - Вычислительная система со множественным потоком команд и множественным потоком данных.

Natural language processing (NLP). Набор заимствованных из информатики и лингвистики методик распознавания естественного языка человека.

Network analysis. Набор методик анализа связей между узлами в сетях. Применительно к социальным сетям позволяет анализировать взаимосвязи между отдельными пользователями, компаниями, сообществами и т.п.

Optimization. Набор численных методов для редизайна сложных систем и процессов для улучшения одного или нескольких показателей. Помогает в принятии стратегических решений, например, состава выводимой на рынок продуктовой линейки, проведении инвестиционного анализа и проч.

Patternre cognition. Набор методик с элементами самообучения для предсказания поведенческой модели потребителей. Predictive modeling. Набор методик, которые позволяют создать математическую модель наперед заданного вероятного сценария развития событий. Например, анализ базы данных CRM-системы на предмет возможных условий, которые подтолкнут абоненты сменить провайдера.

Regression. Набор статистических методов для выявления закономерности между изменением зависимой переменной и одной или несколькими независимыми. Часто применяется для прогнозирования и предсказаний.

Sentimentan alysis. В основе методик оценки настроений потребителей лежат технологии распознавания естественного языка человека. Они позволяют вычленить из общего информационного потока сообщения, связанные с интересующим предметом (например, потребительским продуктом). Далее оценить полярность суждения (позитивное или негативное), степень эмоциональности и проч.

Signal processing. Заимствованный из радиотехники набор методик, который преследует цель распознавания сигнала на фоне шума и его дальнейшего анализа. Spatialan alysis. Набор отчасти заимствованных из геоинформатики и геостатистики [17] методик анализа пространственных данных – топологии местности, географических координат, геометрии объектов. Источником больших данных в этом случае часто выступают геоинформационные системы (ГИС).

Statistics. Наука о сборе, организации и интерпретации данных, включая разработку опросников и проведение экспериментов. Статистические методы часто применяются для оценочных суждений о взаимосвязях между теми или иными событиями.

Supervised learning. Набор основанных на технологиях машинного обучения методик, которые позволяют выявить функциональные взаимосвязи в анализируемых массивах данных. Simulation. Моделирование поведения сложных систем часто используется для прогнозирования, предсказания и проработки различных сценариев при планировании.

Time seriesan alysis. Набор заимствованных из статистики, цифровой обработки сигналов и темпоральной логики методов анализа повторяющихся с течением времени последовательностей данных. Unsupervised learning. Набор основанных на технологиях машинного обучения методик, которые позволяют выявить скрытые функциональные взаимосвязи в анализируемых массивах данных. Имеет общие черты с ClusterAnalysis.

Visualization. Методы графического представления результатов анализа больших данных в виде диаграмм или анимированных изображений для упрощения интерпретации облегчения понимания полученных результатов. Основан на применении методов компьютерной когнитивной графики.

Заключение

Анализ данных больших объемов требует привлечения технологий и средств реализации высоко производительных вычислений. Основными факторами проблемы являются в первую очередь сложность и во вторую физический объем информационной коллекции. Следует отметить, что реальная обработка данных включает еще построение алгоритма и время на его описание и отладку. Уникальные коллекции данных требуют разработки уникальных алгоритмов, что на порядки увеличивает общее время обработки.

Большие объемы данных порождают проблемы при формировании информационных ресурсов из таких данных [18]. По существу большие данные являются новой формой информационного барьера [19]. Большие данные в современном обществе являются частью картины мира [20-23]. Кроме того, эта проблема обу-

славливают постановку и решение новых задач. Это обуславливает развитие интегрированных и комплексных систем и технологий. Повышенное внимание к «большим данным» со стороны «не ИТ специалистов» обусловлено отсутствием практики преодоления информационных барьеров и рассмотрением этого явления как совершенно нового, в то время как оно периодически появляется в развитии человечества и «новым» является не само явление, а новой формой известного явления. С научной точки зрения решение проблемы «большие данные» способствует развитию познанию окружающего мира и построению его целостной картины [20-23].

ЛИТЕРАТУРА

1. Черняк Л. Большие данные–новая теория и практика //Открытые системы. СУБД – 2011. – №10. – С.18-25.
2. Jacobs,A. The pathologies of big data //Communications of theACM. – 2009. – V. 52. – №. 8. – p.36-44.
3. 3.Tsvetkov V. Ya. Complexity Index // European Journal of Technology and Design, 2013, Vol.(1), № 1, p.64-69.
4. LynchC. Bigdata: How do your data grow? //Nature. – 2008. – V. 455. – №. 7209. – p.28-29.
5. Цветков В.Я. Методы и системы обработки и представления видеонформации. – М.:ГКНТ, ВНТИЦентр, 1991. – 113 с.
6. Денисов А.А. Информационное поле. – СПб.: Изд-во "Омега", 1998. – 64 с.
7. Цветков В. Я.Естественное и искусственное информационное поле// Международный журнал прикладных и фундаментальных исследований. – 2014. – №5, ч.2. – С. 178–180.
8. The Fourth Paradigm: Data-Intensive Scientific Discovery, 2009, URL: http://research.microsoft.com/enus/collaboration/fourthparadigm
9. LoekEssers: CERN pushes storage limits as it probes secrets of universe, URL: http://news.idg.no/cw/art.cfm?id=FF726AD5-1A64-6A71-CE987454D9028BDF.
10. Tsvetkov V. Ya. Information Management of Mobile Object // European Journal of Economic Studies, 2012, Vol.(1), №1. P. 40-4
11. Tsvetkov V.Ya. Cognitive information models. // Life Science Journal. 2014; 11(4). pp. 468-471.
12. Кулагин В.П. Проблемы параллельных вычислений // Перспективы науки и образования. 2016. №1. С.7-11.
13. Tsvetkov V. Ya. Information Constructions // European Journal of Technology and Design, 2014, Vol (5), № 3. p.147-152.
14. Tsvetkov V. Ya. Information interaction // European Researcher, 2013, Vol.(62), № 11-1. p.2573.
15. Прорывная технология машинного перевода и вокруг нее. PC WEEK, №9, 12 апреля 2011.
16. Цветков В. Я. Когнитивная кластеризация // Славянский форум, 2016. – № 1(11). – С.233-240.
17. Цветков В.Я. Геостатистика // Известия высших учебных заведений. Геодезия и аэрофотосъемка. – 2007. – №3. – С.174-184.
18. Herodotou H. et al. Starfish: A Self-tuning System for Big Data Analytics // CIDR. – 2011. – Т. 11. – p.261-272.
19. T.A. Ozhereleva. Information Barriers. // European Journal of Technology and Design, 2016, Vol.(11), Is. 1, pp.30-34. DOI: 10.13187/ejtd.2016.11.30 www.ejournal4.com
20. TsvetkovV.Ya. Worldview Model as the Result of Education // World Applied Sciences Journal. – 2014. – 31 (2). – p.211-215.
21. Степин В.С., Кузнецова Л.Ф. Научная картина мира в культуре техногенной цивилизации. М.: Directmedia, 2013.
22. Клягин Н. Современная научная картина мира. М.: Litres, 2016.
23. Вонсовский С. В. Современная естественно-научная картина мира. Екатеринбург: Изд-во Гуманитарного ун-та, 2005.

Информация об авторе
Чехарин Евгений Евгеньевич
(Россия, Москва)
Заместитель начальника центра информатизации
Старший преподаватель кафедры интегрированных информационных систем Института информационных технологий
Московский технологический университет (МИРЭА)
E-mail: tchekharin@mirea.ru

Information about the author
Chekharin Evgeniy Evgen'evich
(Russia, Moscow)
Deputy head of center of information
Senior lecturer of the Department of integrated information systems Institute of information technologies
Moscow Technical University (MIREA)
E-mail: tchekharin@mirea.ru