УДК 7237/4

НУРЖАНОВА С.А., САГЫМБЕКОВ С.Э.
КНУ им. Ж. Баласагына, Бишкек
NURZHANOVA S.A., SAGYMBEKOV S.E.
J. Balasagyn KNU, Bishkek

АНАЛИЗ ИСПОЛЬЗОВАНИЯ «BIG DATA» И ПРОГРАММНЫХ СРЕДСТВ ПОДДЕРЖКИ

Big data" жана программалык колдоо каражаттарын анализдөө

Analysis use of “big data «and software support»

Аннотация: *В статье раскрыто понятие термина «Big Data», проведен анализ «больших данных», его методов и программных средств поддержки, также выявлены особенности применения и роль «Big Data» в современном обществе.*

Аннотация: *Бул беренеде "Big Data" түшүнүгү ачып берилди, «чон маалыматтарды» жана анын ыкмалары, программалык камсыздары, ошондой эле өтүнмөнү карап чыгуунун өзгөчөлүктөрү аныкталган жана заманбап коомго "Big Data" ролу да аныкталды.*

Annotation: *The article fully discloses the concept of the term “Big Data”, analyzes the “big data”, its methods and software support, also reveals the features of the application and the role of “Big Data” in modern society.*

Ключевые слова*: Исследование, анализ, Big Data, Hadoop, MapReduce, Data Mining, краудсорсинг*

Урунттуу сөздөр*: Изилдөө, анализ, Big Data, Hadoop, MapReduce, Data Mining, краудсорсинг*

Keywords: *Research, analysis, Big Data, Hadoop, MapReduce, Data Mining, Crowdsourcing*

В реальном времени размеры инфы вырастают по экспоненциальному закону. Дабы получить конкурентноспособные выдающиеся качества, скорее откликаться на конфигурации, увеличить эффективность изготовления надо добыть, обработать и изучить большую численность данных. И речь идет не о гб и тб данных, с которыми на этот момент имеется возможность преодолеть средний ПК, а о петабайтах и эксабайтах. Для работы с этими размерами инфы инженерам довелось модернизировать инструменты для анализа всех данных. Например, в 2000-х годах развилось понятие «BigData» (Большие Данные). Технологии большущих данных дают возможность обработать большущий размер неструктурированных данных, классифицировать их, изучить и обнаружить закономерности там, где человечий мозг ни разу бы их не обнаружил. Это раскрывает абсолютно свежие способности по применению данных.

Понятие термина «Big Data» (Большие данные)

Гигантские данные— обозначение структурированных и неструктурированных данных больших объёмов и значимого обилия, действенно обрабатываемых горизонтальную масштабируемыми программными инструментами, показавшимся в конце 2000-х годов и другими классическими системами управления базами данных и заключениям класса Business Intelligence. [11]

В качестве определяющих данных для большущих данных обычно выделяют «три V»: объём (англ. volume, в значении величины физиологического объёма), скорость (velocity в смыслах как скорости прироста, например, и надобности скоростной обработки и получения результатов), разнообразие (variety, в значении способности одновременной обработки всевозможных типов, структурированных и полуструктурированных данных); в последующем появились всевозможные варианты и интерпретации сего симптома [3]

Принципы работы «Big Data».

Техники и методы анализа, применимые к Big data по McKinsey[4]

- методы класса Data Mining: изучение ассоциативным правилам, классификация (методы категоризации свежих данных на базе основ, раньше применённых к уже наличествующим данным), кластерный тест, регрессионный анализ;

краудсорсинг — категоризация и обогащение данных силами широкого, неопределённого круга лиц, привлечённых на основании общественной оферты, без введения в трудовые отношения; - смешение и

интеграция данных — комплект техник, позволяющих интегрировать разнородные данные из всевозможных источников для способности глубинного анализа, в качестве примеров этих техник, элементах данный класс способов, приводятся цифровая

247

обработка сигналов и обработка натурального языка (включая тональный анализ);

- машинное изучение, охватывая изучение с учителем и без учителя — внедрение моделей, построенных на основе статистического анализа или же машинного изучения для получения всеохватывающих мониторингов на базе базисных моделей;
- искусственные нейронные сети, сетевой тест, оптимизация, в что количестве генетические алгоритмы;
- распознавание образов; - прогнозная аналитика; имитационное моделирование;
-пространственный тест — класс способов, использующих топологическую, геометрическую и географическую информацию в данных;
- статистический тест, в качестве примеров способов приводятся рядов;
- визуализация аналитических данных — представление инфы в облике рисунков, диаграмм, с внедрением интерактивных вероятностей и анимации как для получения итогов, например и для применения в качестве начальных данных для последующего анализа.

Программные средства поддержки – BIGDATA

Hadoop считается одной из основных технологий BigData [11]. Разработка была иниицирована в начале 2005 года Дугом Каттингом (Doug Cutting) с целью возведения программной инфраструктуры распределенных вычислений для плана Nutch – свободной программной поисковой машины JAVA, ее идеологической почвой стала объявление служащих Гугл Джеффри Дина и Санжая Гемавата о вычислительной концепции MapReduce. Свежий план был назван в честь игрушечного слоненка ребенка основоположника плана. Разработка Hadoop дает собой программный фреймворк, позволяющий беречь и обрабатывать данные с поддержкой компьютерных кластеров, применяя парадигму MapReduce. MapReduce – это фреймворк для вычисления кое-каких наборов распределенных задач с внедрением большущего числа компов (называемых “нодами”), образующих кластер. Работа MapReduce произведено из двух шагов: Map и Reduce. На Map шаге случается предшествующая обработка входных данных. Для сего раз из компов (называемый ключевым узлом mastermode) получает входные данные задачки, разграничивает их на части и передает иным компам (рабочим узлам - workemode) для подготовительной обработки. Заглавие этот шаг получил от похожей функции высочайшего около [12].

На Reduce – шаге случается свертка сначала обработанных данных. Ключевой узел получает ответы от трудящихся узлов и на их базе создает итог – заключение задачки, которая в начале формулировалась.

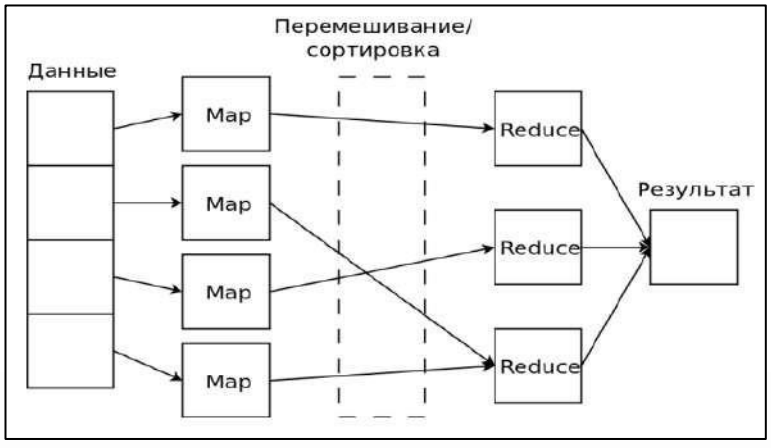

Рис 2. Схема вычислений Map Reduce

Для анализа данных используются всевозможные способы. Проанализируем главные из них:

1. Методы класса DataMining (глубинный тест данных). Главная индивидуальность DataMining – это хитросплетение широкого математического инвентаря (от традиционного статистического анализа до свежих кибернетических методов) и последних достижений в сфере информационных технологий [8].

2. A/B-тестирование (A/Btesting, Splitesting) – способ рекламного изучения, сущность которого заключается в том, собственно что контрольная группа составляющих сравнивается с набором тестовых групп, в несколько раз или же некоторое количество характеристик были изменены, для такого, дабы узнать, какие из перемен делают лучше мотивированный показатель. Гигантские данные дают возможность выполнить большую численность операций и этим образом получить статистически надежный результат.

3. Краудсорсинг – способ сбора данных из большущего числа источников.

4. Машинное изучение. Назначение в информатике (исторически за ним зафиксировалось заглавие «искусственный интеллект»), которое преследует задача сотворения алгоритмов самообучения на базе анализа эмпирических данных и др. [9]. Есть большое количество инструментов BigData. Разглядим известные из них: NoSQL (notonly SQL, не лишь только SQL), в информатике – термин, обозначающий ряд раскладов, нацеленных на реализацию хранилищ баз данных, имеющих немаловажные отличия от моделей, применяемых в классических реляционных СУБД с доступом к сведениям способами языка SQL. Классические СУБД определяются на запросы ACID к транзакционной системе: атомарность (atomicity), согласованность (consistence), обособленность (isolation), надежность (durability), за это время как в NoSQL взамен ACID имеет возможность рассматриваться комплект качеств BASE:

- Базисная доступность (basoc availability) – любой запрос гарантированно заканчивается (успешно или же безуспешно).

- Гибкое положение (softstate) – положение системы имеет возможность переменяться с периодом, в том числе и без ввода свежих данных, для заслуги согласования данных.

- Согласованность в конечном (eventualconsistency) – данные имеют все шансы быть коекакое время рассогласованы, но приходят к согласованию сквозь кое-какое время.

Термин BASEбыл предложен Эриком Брюэром, создателем аксиомы CAP, сообразно которой в распределенных вычислениях возможно гарантировать лишь только два из трех качеств:

Кластеграмма

Способ визуализации, применяемый при кластерном анализе. Демонстрирует, как отдельные составляющие большого количества данных соотносятся с кластерами по мере конфигурации их числа. Выбор рационального числа кластеров – значимый элемент кластерного анализа.

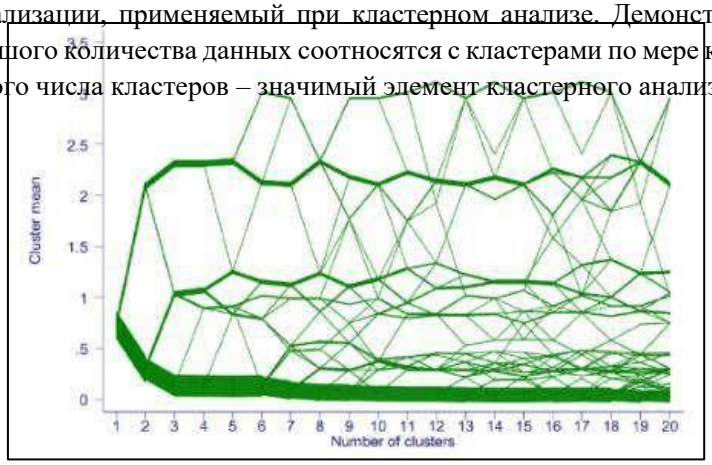

Рис. 3. Метод класстерного анализа

Исторический поток

Может помочь наблюдать за эволюцией документа над созданием которого трудится в одно и то же время большущая численность создателей. В частности, это обычная обстановка для сервисов wiki и вебсайта tadviser в количестве. По горизонтальной оси отменятеся время, по вертикальной – лепта всякого из соавторов, т.е. размер введенного слова. Любому оригинальному создателю присваивается конкретная краска на диаграмме. Приведенная диаграмма – итог анализа для текста «ислам» в Википедии. Отлично видать, как росла энергичность создателей с течением времени

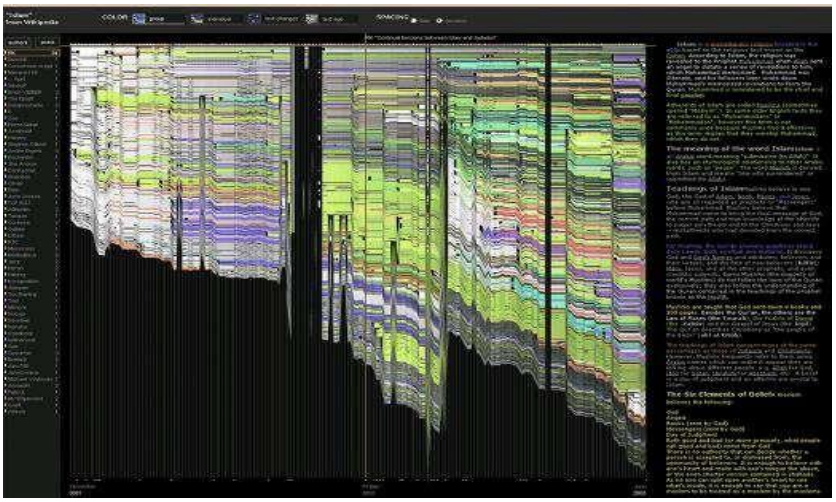

Рис. 4. результат анализа для слова «ислам» в Википедии

Пространственный поток

Данная диаграмма разрешает выслеживать пространственное рассредотачивание инфы. Приведенная в качестве примера диаграмма построена с поддержкой обслуживания New York Talk Exchange. Она визуализирует напряженность обмена IP-трафиком меж Нью-Йорком и другими городками мира. Чем ярче трасса – что более данных передается за единицу времени. Этим просто, не оформляет труда отметить ареалы, более ближайшие к Нью-Йорку в контексте информационного обмена.

Рис. 5. Отслеживание пространственного распределения информации

Особенности использования и роль «Big Data» в современном обществе

Исследуя разнообразие передовых технологий сбережения и обработки данных, появляется логичный вопрос. Для чего выдуманы способы и расклады, именуемые «Big Data» ? Собственно что в данном оригинального, как возможно применить информацию, обработанную с поддержкой данных технологий и отчего фирмы готовы инвестировать в становление большущих данных большие средства ?

Для начала, в различие от Big Data , обыденные базы данных (БД), не имеют все шансы беречь и возделывать эти большие размеры данных (сотни и тыс. терабайт). И речь в том числе и не об специалисте, а лишь только о сбережении данных .

В традиционном осознании БД предопределена для резвой обработки (хранение, изменение) сравнительно маленьких размеров данных или же для работы с большущим потоком записей маленького объема, т. е. транзакционная система. С поддержкой Big Data как один принимается решение данная главная задачка – успешное сбережение и обработка большущих размеров данных . [8] Во-2-х, в Big Data структурируются разнотипные сведения, которые поступают из всевозможных источников (изображения, фото, видео, аудио и текстовые документы) в раз единственный, понятный и приемлемый для последующей работы вид. В-третьих, Big Data случается составление специалисты и возведение четких мониторингов на основании приобретенной и обработанной инфы. Для чего это надо и где имеет возможность быть использовано на практике ? Для наглядности и для такого, дабы сконструировать ответ ординарными текстами, разглядим на случае обычных бизнес-задач в маркетинге. Владея подобной информацией, как:

-- абсолютное осознание о собственной фирмы и коммерциале, в что количестве с точки зрения статической инфы и цифр;

-- досконалные данные о конкурентах;

-- свежая и досконалная информация о собственных клиентах;

-- все это дозволит преуспеть в вербовании свежих покупателей, важно увеличить степень предоставляемого обслуживания текущим покупателям, чем какого-либо другого взять в счет доминирования над ними. Беря во внимание перечисленные выше итоги, коих разрешает добиться Big Data, и разъясняет влечеиие фирм, пытающихся достичь базар, вкладываться в эти современные способы обработки данных сейчас, дабы получить наращивание продаж и сокращение потерь на следующий день. А в случае, если больше непосредственно, то:

-- наращивание добавочных продаж и кросс продаж за счет наилучшего познания предпочтений клиентов;

-- разведка известных продуктов и оснований – отчего их приобретают или же наоборот;

-- улучшение предоставляемой предложения или же продукта;

-- увеличение свойства сервиса клиентов;

-- увеличение преданности и клиенто-ориентированности;

-- предупреждение афер (больше животрепещуще для банковской сферы); --

понижение бесполезных затрат .

Раз из более приятных и известных на нынешний день примеров, о котором возможно прочесть во множестве источниках сети Онлайн, связан с фирмой Apple, которая собирает данные о собственных юзерах с поддержкой выполняемых приборов: телефонный аппарат, планшет, часы, компьютер. Как раз по причине присутствия подобный системы компания обладает большой численностью инфы о собственных юзерах и в последующем пользуется ее для получения выгоды. И аналогичных примеров на нынешний денек возможно отыскать единое большое количество.

Заключение

Технологии большущих данных дают возможность обработать большущий размер неструктурированных данных, классифицировать их, изучить и обнаружить закономерности там, где человеческий мозг ни разу бы их не обнаружил. Это раскрывает абсолютно свежие способности по применению данных. Не обращая внимания на очевидные выдающиеся качества и плюсы BigData, есть и собственные трудности. Технологии BigData довольное бодрое содержание. Почти все считают, собственно что у BigData большущее будущее и это аутентичный прорыв в информационных разработках. И правду, у технологий большущих данных большая область использования и в любом порознь взятом случае возможно извлечь пользу от применения данных технологий. Но надо внятно воспринимать плюсы и минусы. Технологии большущих данных начинают внедряться во все ветви нашей жизни: деньги, здравоохранение, сельское хозяйство, телекоммуникации, розничная торговля, воспитание, городское управление, ЖКХ, военнослужащая индустрия, госструктуры и т.д.

Гигантская доля ИТ фирм организуют или же поддерживает прохождение работниками курсов повышения квалификации. При проведении выборочного опроса, все ИТ фирмы замечали, собственно, что желают расширения штата служащих. Главы фирм, считают что на все направлен ощущается недостаток сотрудников в их компаниях. На вопрос: «С чем, по Вашему мнению, связан недостаток сотрудников в данных направлениях?» они подчеркнули, что невысокий степень подготовки выпускников Институтов и трудовую миграцию. На вопрос
«Какие кадры станут необходимы вашей фирмы и в целом рынку КР в наиближайшие 3года?» они считают, что в наиближайшие 3 года все еще станут популярными зн атоки веб - и мобильной разработки, впрочем, беря во внимание проворно меняющиеся технологии в ИТ, в наиближайшие годы Кыргызстан еще не станет брать знатоков высочайшего значения в обл астях: искусственного происхождения разума, урок о большущих данных (big data science). Беря во внимание, собственно что основная масса ИТ-компаний аутсорят собственные предложения за границу, важным аспектом для них остается знание работниками английского языка. Главы еще дали совета по таким направлениям, какие нужно подключить в образовательную программку ИТнаправлений. На этот момент невозможно буквально квалифицировать будущее BigData, но эти технологии, имеют ряд неопровержимых превосходства. Гигантские данные раскрывают перед нами свежие горизонты в планировании, образования, здравоохранения и иных секторах экономики. В случае, если их становление станет делиться, то технологии большущих данных имеют все шансы поднять информацию, как момент изготовления, на абсолютно свежую высококачественную ступень. Информация будет не только равноценна труду и состоянию, но и вполне вероятно будет наиважнейшим ресурсом прогрессивной экономики.

Технологии BigData благополучно реализуются в промышленности. В инфографике отражены главеные покупатели: банки, телеком, ритейл, энергетика, медицина и управление городской инфраструктурой.

Список цитируемых источников

Нормативно-правовая литература

1. Распоряжение Правительства Кыргызской Республики от 14 октября 2016 года № 436-р **Специальная литература**

1. Корнев М.С., История понятия «большие данные» (Big Data): словари, научная и деловая периодика. УДК 070:004.6 (Дата обращения)
2. Черняк Л. Большие Данные – новая теория и практика // Открытые системы. СУБД. 2011. № 10. С. 18–25; цит. по: URL: https://www.osp.ru/ os/2011/10/13010990/ (дата обращения: 05.07.2019).
3. 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.

4. Академия BIG DATA: Введение в аналитику больших массивов данных: Информация // Национальный Открытый Университет «ИНТУИТ». URL: https://www.intuit.ru/studies/courses/12385/1181/info (дата обращения: 30.04.2019).

5. Аналитический обзор рынка Big Data // Хабр. URL: https://habr.com/company/moex/blog/256747/ (Дата обращения: 4.04.2019).

6. Streamline Your Big Data Platform // ORACLE. URL: https://www.oracle.com/big-data/index.html (Дата обращения: 30.04.2019) MapReduce and Teradata Aster SQL-

7. MapReduce // Teradata. URL: https://www.teradata.com/products-and-services/Teradata-Aster/teradata-aster-sqlmapreduce (Дата обращения: 30.03.2019)

8. Коновалов М. В. Big Data. Особенности и роль в современном бизнесе [Текст] // Технические науки: проблемы и перспективы: материалы VI Междунар. науч. конф. (г. Санкт-Петербург, июль 2018 г.). — СПб.: Свое издательство, 2018. — С. 8-10. — URL https://moluch.ru/conf/tech/archive/288/14418.

9. Сухобоков А. А., Лахвич Д. С., Влияние инструментария BigData на развитие научных дисциплин, связанных с моделированием // Наука и Образование. МГТУ им. Н.Э. Баумана. Электрон.журн. 2015. № 03. С. 207– 240. (дата обращения 07.05.2019).

Интернет-источники

1. Как большие данные стали одной из самых интересных задач IT-индустрии, https://postnauka.ru/specials/bigdata (дата обращения 19.05.2019)

2. Что такое Big Data? https://postnauka.ru/faq/46974 (Дата обращения 19.05.2019)

3. Большие данные. Материал из Википедии — свободной энциклопедии. https://ru.wikipedia.org (Дата обращения 19.05.2019)

4. Big Data. Материал из Национальной библиотеки им. Н. Э. Баумана. https://ru.bmstu.wiki/Big_Data (Дата обращения 4.04.2019)

5. История Big Data восходит к практикам общественного порядка XIX века. https://22century.ru/popularscience-publications/big-data-problems (Дата обращения 19.05.2019)

6. Большие данные // Википедия: сайт. Режим доступа: https://ru.wikipedia.org/wiki/Большие_ данные (дата обращения 13.05.2019)

7. BigData в российских банках. Начало большого пути // PCWEEK: сайт Режим доступа: http://www.pcweek.ru/idea/article/detail.php?ID=176526 (дата обращения 07.05.2019).

8. Особенности, методы и стадии DataMining в сфере информационных технологий // Веб студия X-ON: сайт. Режим доступа: http://old.x-on.ru/work/2-2-metody-i-stadii-data-mining/ (дата обращения 08.05.2019)

9. Предсказательная аналитика // Википедия: сайт. Режим доступа: https://ru.wikipedia.org/wiki/ Предсказательная_аналитика (дата обращения 07.04.2019).

10. Теория распознавания образов // Википедия: сайт. Режим доступа: https://ru.wikipedia.org/wiki/ Теория_распознавания_образов (дата обращения 17.04.2019).

11. Apache Hadoop // Apache Hadoop: сайт. Режим доступа: http://hadoop.apache.org (дата обращения 4.05.2019).

12. MapReduce // Википедия: сайт. Режим доступа: https://ru.wikipedia.org/wiki/MapReduce (дата обращения 17.05.2019).

13. Module 2: The Hadoop Distributed File System//Yahoo! Developernetwork: сайт. Режим доступа: https://developer.yahoo.com/hadoop/tutorial/module2.html (дата обращения 08.05.2019).

14. Примеры использования Big Data. https://www.xelent.ru/blog/primeryi-ispolzovaniya-big-data-ch-2/ (дата обращения 20.05.2019)

***Рецензент: Валеева А.А.**- кандидат физико-математических наук, профессор КГТУ им.И. Раззакова*