Михнев Илья Павлович
канд. техн. наук, доцент, доцент,
Заслуженный работник науки и образования
Новикова Анастасия Александровна
студентка
Петросян Мелине Кареновна
студентка
Волгоградский институт управления (филиал)
ФГБОУ ВО «Российская академия народного хозяйства и
государственной службы при Президенте РФ»
г. Волгоград, Волгоградская область

DOI 10.21661/r-470260

БОЛЬШИЕ ДАННЫЕ (BIG DATA) И НОВЫЕ ТЕХНОЛОГИИ БУДУЩЕГО ДЛЯ ОБРАБОТКИ ГЛОБАЛЬНОЙ ИНФОРМАЦИИ

Аннотация*: в статье рассматриваются новые технологии будущего Big Data для обработки, хранения и использования больших данных. Изложены методы обработки неструктурированной информации, серия подходов и инструментарий больших данных. Представлены современное состояние и тенденции развития технологий Big Data.*

Ключевые слова*: инструментарий больших данных, СУБД, базы данных, большие данные, эксабайтная информация, методы анализа Big Data.*

Большинство из нас, являясь пользователями современной информации, хотя бы раз слышали крылатые фразы: «Кто владеет информацией, тот владеет миром» и «Кто управляет информацией, тот управляет всем миром». Полученная раньше других, актуальная информация, в свое время дала возможность Ротшильдам вести беспроигрышную игру на бирже. Ротшильды не только придумали вышеупомянутую фразу, они сделали всё, чтобы нужная информация попала к ним в первую очередь. Глобальная информация всегда обладала исключительной важностью. С помощью нужной информации многие становились

миллиардерами, свергали неугодные правительства и переворачивали целые государства, т.е. глобальная информация всегда имела масштабную силу и власть. В современном мире глобальные объемы информации увеличиваются по экспоненциальному закону. Чтобы максимально быстро рефлектировать на современные изменения, получать преимущества над конкурентами и повышать эффективность производства, необходимо собирать, анализировать и обрабатывать огромное количество неструктурированных данных. Имеется в виду не гигабайты (2^{30} байт) и терабайты (2^{40} байт) традиционных данных, которые на современном этапе может обрабатывать обычный ПК, а петабайты (2^{50} байт), эксабайты (2^{60} байт), зеттабайты (2^{70} байт) и йоттабайты (2^{80} байт) неопределённо структурированных данных. Для обработки таких объемов информации необходимо модернизировать инструментарий для анализа всех данных или задействовать квантовые компьютеры [1; 3–6].

В современном мире, одним из ключевых драйверов развития информационных технологий, являются «Большие данные» (Big Data). В сущности, понятие «Большие данные» подразумевает обработку информации разнообразного состава и огромного объема, очень быстро обновляемой, находящейся в различных источниках для увеличения эффективности работы, создания новых продуктов и повышения конкурентной способности [2–3; 11].

Редактор журнала Nature Линч Клиффорд, подготовил специальный номер журнала на тему: «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?». В этом журнале были изложены исследования о феномене взрывного роста объёмов и многообразия обрабатываемых данных и технологических перспективах в парадигме вероятного скачка «от количества к качеству». Именно тогда и появился сам термин Big Data, то есть 3 сентября 2008 года. Термин был предложен по аналогии с расхожими в деловой англоязычной среде метафорами «Большая нефть», «Большая руда» [4–7; 10].

Big Data объединяют технические средства и новейшие технологии, которые находят смысл из различных данных на сверхэкстремальном пределе

практичности. Big Data требует специальных подходов, инструментария и методов, которые значительно отличаются от традиционных классических. Новые технологии будущего на основе Big Data – это целая серия подходов, инструментария и методов обработки значительного многообразия неструктурированных данных огромных объёмов для получения воспринимаемых пользователем новых результатов. Также эффективных в условиях непрерывного роста, дислокация по разнообразным узлам глобальной сети, альтернативных традиционным СУБД и решениям класса Business Intelligence. В данную серию подходов включают новейшие средства массовой одновременной обработки неструктурированных больших данных, прежде всего, возможностями таких категорий, как NoSQL, алгоритмами MapReduce, спецпрограммными контурами и Hadoop библиотеками. В качестве главных характеристик для Big Data выделяют: velocity, variety, volume, то есть скорость, многообразие и физический объём. Под скоростью понимают скорость прироста, высокоскоростную обработку для получения необходимых результатов, а под многообразием – возможность одновременно обрабатывать различные типы данных (полностью структурных и неструктурных). В истинном понимании термина Big Data, в действительности, только очень крупные организации обладают большими данными, так как даже десятки террабайт собранной информации таковыми просто не являются. Когда традиционных подходов и стандартных решений уже не хватает, террабайтная база данных (БД) реляционного типа – это DB-Highload, но не как не Big Data. Разница между этими понятиями заключается в возможности строить гибкие запросы. Обычные реляционные БД подходят для достаточно быстрых и однотипных запросов, а на сложных и гибко построенных запросах нагрузка просто превышает все разумные пределы и использование СУБД становится неэффективным. При этом, методы анализа Big Data вполне применимы и к данным, которые изначально большими не являются. Более того, аналитика, построенная на статистическом анализе и машинном обучении может быть полезна во многих проектах. К новейшим Big Data применимы следующие технологии анализа и методы обработки неструктурированных данных [5; 7]:

– глубинный анализ данных по методу Data Mining. Основная особенность Data Mining заключается в сочетании широкого математического инструментария (от традиционного анализа статистики до новых кибернетических подходов) и последних разработок в информационной сфере. К алгоритмам и спецметодам Data Mining можно отнести метод ограниченного перебора, искусственные нейронные сети, эволюционное программирование и генетические алгоритмы, правила символьной основы, различные методы визуализации данных, методы ближайшего соседа, методы опорных векторов, байесовские сети, линейная регрессия и корреляционно-регрессионный анализ. А также допустимы иерархические и неиерархические методы кластерного анализа, в том числе алгоритмы k-средних и k-медианы, методы поиска ассоциативных правил, в том числе алгоритм Apriori [6; 9–11];

– метод маркетингового исследования (A/Btesting, Splittesting). В этом методе контрольная группа элементов сравнивается с заданным набором тестовых групп. Причем, в наборе групп один или несколько показателей изменяют для того, чтобы выяснить, какие из изменений улучшают итоговый показатель. При этом Big Data позволяют провести огромное количество итераций и таким образом получить достоверный результат с точки зрения статистики;

– современная методика сбора данных из огромного количества источников информации («Crowdsourcing» – Краудсорсинг);

– «искусственный интеллект» или машинное обучение для создания алгоритмов самообучения на основе общего анализа эмпирических данных;

– набор методик анализа существующих связей между различными узлами в глобальных сетях – «Сетевой глобальный анализ». Набор методик в социальных сетях может анализировать взаимные связи между конкретными пользователями, отдельными компаниями, различными сообществами и т. д.;

– целый кластер методов анализа больших данных, концентрирующийся на прогнозировании будущего поведения разных объектов и субъектов для принятия оптимальных решений, т.е. «Прогнозная аналитика»;

– метод для построения модели, описывающей процессы прохождения как бы в действительности, т.е. «Имитационное моделирование». Подобную модель можно использовать как для единичного испытания, так и для необходимого множества (результаты при этом, будут определяться случайным характером процессов) с достаточно устойчивой статистикой;

– распознавание образов – Раздел информационных технологий и смежных дисциплин, образовывающий основополагающие методы аутентификации и идентификации процессов, объектов, явлений, и т. д., которые характеризуются итоговым набором признаков и свойств.

Технологии Big Data используют большое множество инструментария, самыми популярными из них являются: NoSQL (Not only Structured Query Language – SQL, не только «язык структурированных запросов»). То есть инструмент, с большим рядом подходов, направленный на реализацию хранилищ БД и БнД, имеющих значительные отличия от традиционных моделей (используемых в реляционных СУБД с обеспечением языка SQL).

Современные СУБД строятся на требовании ACID к транзакционной системе. А именно атомарности (Atomicity), согласованности (Consistency), изолированности (Isolation), надёжности (Durability), тогда как в NoSQL вместо ACID рассматривает набор свойств BASE:

– базовая доступность (Basic Availability), каждый запрос успешно или безуспешно, но гарантированно завершается;

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

– согласованность, в конечном счёте (Eventualconsistency), т.е. данные некоторое время могут быть несогласованны, но будут согласованы через заданное время.

Термин «BASE» был предложен Эриком Брюером, автором теоремы CAP, согласно которой в распределённых вычислениях можно обеспечить только два из трёх свойств: согласованность данных, доступность или устойчивость к разделению. Подобные системы на основе BASE не могут использоваться в любых

приложениях: для функционирования биржевых и банковских систем использование транзакций является необходимостью. В то же время, свойства ACID, какими бы желанными они ни были, практически невозможно обеспечить в системах с многомиллионной веб-аудиторией, вроде amazon.com. Следовательно, инженеры-проектировщики систем NoSQL жертвуют согласованностью данных ради достижения двух других свойств из теоремы CAP. Решения систем NoSQL отличаются не только проектированием с учётом масштабирования, но и другими характерными NoSQL-решениями [1; 7; 11]:

– использование разных типов хранилищ;

– разработка БД с возможностью «без задания схемы»;

– применение множества процессоров (многопроцессорность);

– для увеличения производительности используется «Линейная масштабируемость»;

– «инновационность NoSQL» позволяет открыть множество возможностей для обработки и хранения больших данных;

– скорость и сокращение времени разработки. Даже при минимальном объеме данных у пользователей есть возможность оценить уменьшение времени отклика системы с сотен миллисекунд до миллисекунд.

Одной из основополагающих технологий Big Data является Hadoop. Инициирование данной разработки было в 2005 году Дугом Каттингом. Цель – построение программной инфраструктуры распределённых вычислений для проекта Nutch – свободной поисковой машины на Java. Новый проект назвали в честь игрушечного слонёнка, который был у сына основателя проекта. Технология Hadoop является программным «фреймворком», позволяющим хранить и обрабатывать большие данные с помощью кластеров компьютера, используя модель MapReduce – фреймворк для вычисления наборов распределенных задач с использованием огромного количества ПК (называемых «нодами»), образующих кластер. Работа MapReduce состоит из двух шагов: Map и Reduce. На первом шаге осуществляется предварительная обработка входных данных, т.е. один из компьютеров (главный узел – Masternode) получает входные данные задачи,

разделяет их на части и передает другим компьютерам (рабочим узлам – Worker-node) для предварительной обработки. Название первый шаг получил от одноименной функции высшего порядка [10–11]. На втором Reduce-шаге происходит свёртка предварительно обработанных данных. Masternode получает ответы от Workernode и на их основе формирует решение задачи, которая изначально формулировалась – т.е. результат. Такой подход позволяет выстраивать кластер высокой производительности на базе серверов «lowend» или «middle-end», что снижает стоимость решения по сравнению с одним высокопроизводительным сервером. В основе технологии лежит HDFS (Hadoop Distributed File System) – распределенная файловая система, созданная для хранения очень большого объема информации (терабайт или петабайт) и обеспечения высокой скорости доступа к этой информации. Информация хранится в избыточной форме на множестве ПК для обеспечения их устойчивости при возможных ошибках и высокой доступности параллельным приложениям. Если один или несколько узлов кластера выходят из строя, то риск потери информации сводится к минимуму и кластер продолжает работу в штатном режиме.

Список литературы

1. Силен Д. Основы Data Science, Big Data. Python и наука о данных / Д. Силен. – М.: Питер, 2017. – 354 с.

2. Михнев И.П. Технологии Big Data и их применение в сфере современного высшего образования / И.П. Михнев, А.Д. Челнокова, А.Д. Реут // Развитие современного образования: от теории к практике: Материалы IV Междунар. науч.-практ. конф. (Чебоксары, 19 март 2018 г.) / Редкол.: О.Н. Широков [и др.]. – Чебоксары: ЦНС «Интерактив плюс», 2018.

3. Фрэнкс Б. Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики / Б. Фрэнкс. – М.: Альпина Диджитал, 2014. – 370 с.

4. Моррисон А. Большие Данные: как извлечь из них информацию. Технологический прогноз / Ежеквартальный журнал. – 2010. – №3. – С. 22–29.

5. Михнев И.П. Информационная безопасность в современном экономическом образовании // Международный журнал прикладных и фундаментальных исследований. – 2013. – №4. – С. 111–113.

6. Михнев И.П. Обучение и контроль знаний студентов с помощью UniTest // Фундаментальные исследования. – 2008. – №1. – С. 94–95.

7. Банько Ю.А. Современные компьютерные угрозы: что реально угрожает бизнесу? / Ю.А. Банько, А.М. Кокорева, науч. рук. И.П. Михнев // Приоритетные направления развития образования и науки: Материалы IV Междунар. науч.-практ. конф. (Чебоксары, 24 дек. 2017 г.) / Редкол.: О.Н. Широков [и др.] – Чебоксары: ЦНС «Интерактив плюс», 2017. – С. 169–171.

8. Михнев И.П. Информационная безопасность на просторах мобильного интернета // Образовательные ресурсы и технологии. – 2015. – №4 (12). – С. 66–70.

9. Черняк Л. Большие Данные – новая теория и практика // Открытые системы. СУБД. – 2011. – №10. – С. 36–41.

10. Что такое Big data: собрали всё самое важное о больших данных // RUSBASE [Электронный ресурс]. – Режим доступа: https://rb.ru/howto/chto-takoe-big-data/ (дата обращения 11.03.2018).

11. Big Data: проблема, технология, рынок [Электронный ресурс]. – Режим доступа: http://compress.ru/Article.aspx?id=22725 (дата обращения: 10.03.2018).