УДК 004.62 https://doi.org/10.15407/pp2019.03.086

*В.А. Резниченко*

# ЧТО ТАКОЕ BIG DATA

В статье делается попытка раскрытия сути понятия Big Data на основе анализа материалов из различных источников. Даются определяющие характеристики Big Data, приводится их классификация, кратко описывается история возникновения и развития, представлены основополагающие принципы работы, кратко излагаются методы и технологии анализа и визуализации, описывается жизненный цикл управления данными с использованием технологии Big Data.
Ключевые слова: большие данные, технология больших данных, база данных, жизненный цикл.

## Введение

“Так что же такое Big Data? Это неожиданно обрушившаяся на человечество лавина данных, это принципиально новая информационная технология, это также можно считать техническая и технологическая революции в информатике. Показательно, что из более чем 153 миллиона страниц в Web, содержащих словосочетание Big Data, 122 миллиона содержит еще и слово definition – более двух третей пишущих о Big Data пытается дать свое определение. Такая массовая заинтересованность свидетельствует в пользу того, что, скорее всего, в Big Data есть что-то качественно иное, чем то, к чему подталкивает нас обыденное восприятие этого словосочетания. В этой статье делается попытка раскрытия сути понятия Big Data на основе анализа материалов из различных источников.

## 1. Материальные и информационные технологии

Приведем рассуждения относительно материальных и информационных технологий, почерпнутые из [1]. К информационной технологии надо относиться как к материальной технологии. Практически все известные материальные технологии сводятся к процессу переработки, обработки или сборки специфического для них исходного сырья или каких-то иных компонентов с целью получения качественно новых продуктов.

Логически информационные технологии мало чем отличаются от материальных технологий, на входе сырые данные, на выходе – структурированные, в форме, более удобной для восприятия человеком, данные, извлеченная из них информация, которая силой интеллекта (естественного или искусственного) превращается в полезное знание. Данные – это выраженные в разной форме сырые факты, которые сами по себе не несут полезного смысла до тех пор, пока не поставлены в контекст, должным образом не организованы и не упорядочены в процессе обработки. Информация появляется в результате анализа обработанных данных человеком (компьютером), этот анализ придает данным смысл и обеспечивает им потребительские качества.

На информационные технологии должны распространяться общие закономерности, согласно которых развиваются все остальные технологии, а это прежде всего увеличение количества перерабатываемого сырья способствует повышению качества переработки.

## 2. Проблемы больших данных

Мировой объем оцифрованной информации растет по экспоненте. Начиная с 1980-х годов цифровая информация удваивается каждые 40 месяцев. Поданным компании IBS, к 2003 году мир накопил 5 эксабайтов данных (1 ЭБ = 1 млрд гигабайтов), а теперь это количество порождается каждые два дня. К 2008 году этот объем вырос до 0,18 зеттабайта (1 ЗБ = = 1024 эксабайта), к 2011 году – до 1,76 зеттабайта, к 2013 году – до 4,4 зеттабайта. В мае 2015 года глобальное количество данных превысило 6,5 зеттабайта. К 2020 году, по прогнозам, человечество

сформирует 40–44 зеттабайтов информации, а к 2025 г. – 163 зеттабайт. В настоящее время площадь всех крупных датацентров в мире равна площади 6000 футбольных полей Как справиться с такими объемами?

Приведем следующую цитату из [1]: «Данных становится все больше и больше, но при всем этом упускается из виду то обстоятельство, что проблема отнюдь не внешняя, она вызвана не столько обрушившимися в невероятном количестве данными, сколько неспособностью старыми методами справиться с новыми объемами. Наблюдается дисбаланс – способность порождать данные оказалась сильнее, чем способность их перерабатывать.» Под именем Big Data скрывается намечающийся качественный переход в компьютерных технологиях, способный повлечь за собой серьезные изменения. Не случайно этот переход называют новой технической революцией.

## 3. Определяющие характеристики Big Data

Для Big Data были сформулированы определяющие характеристики. Впервые в 2001 г. признаки «Три V» выделил ведущий аналитик Gartner Дуг Лани [2], а именно, объем, скорость, разнообразие.

**Volume (объем).** Считается, что Big Data начинаются с объемов в петабайты (1015 байт). Чтобы представить, что это за объем, приведем пример. В Национальной библиотеке Украины им. В.И. Вернадского функционирует портал Научной периодики Украины. Редакции более 2700 научных периодических изданий Украины предоставляют все свои статьи на протяжении 10 лет. За это время объем портала составил около 1 миллиона статей. Если предположить, что размер статьи в среднем составляет 1 МБ, то объем ресурсов этого портала составляет 1 ТБ. Это на три порядка ниже минимального объема для Big Data, то есть через десять тысяч лет успешного функционирования этого портала он накопит объемы, характерные для Big Data.

Big Data появляются тогда, когда сотни миллионов людей объединяются в сообщества и выкладывают свои информационные ресурсы, либо объединенные центры научных исследований предоставляют данные результатов своих исследований, например в 2017 году дата-центр CERN превысил размер 200 петабайт и ежегодно этот объем увеличивается на 15 петабайт. Если поместить в DVD все порожденные в мире за день данные и положить эти диски друг на друга, то получится стопка, дважды превышающая расстояние до Луны.

**Velocity (скорость).** Является одной из наиболее важных характеристик Big Data с точки зрения их практического использования. Под скоростью подразумевается как скорость прироста (поступления, накопления) данных, так и скорость их обработки с целью получения конечных результатов. Кроме того, в эту категорию включаются характеристики интенсивности и объемов информационных потоков. Для этого технология обработки таких данных должна допускать возможность их анализа уже в момент их порождения (иногда называемой «оперативной аналитикой » - in- memory analytics), то есть до того, как они попадут в хранилище данных. Несколько цифр, характеризующих эту категорию, которые взяты из [3] и некоторых других источников.

**YouTube:** Имеет более 1 миллиарда зарегистрированных пользователей и ежемесячно сайт посещают 1,9 миллиарда пользователей. Ежеминутно закачивается новых фильмов на 100 часов и скачивается фильмов на 700 тысяч часов. Для просмотра фильмов, выгруженных в YouTube в течении дня, потребуется 15 лет.

**Facebook:** Имеет 1,4 миллиарда пользователей. Ежедневно на сайт выгружается 100 терабайт данных и ежеминутно ставятся более 34 тысячи лайков. Каждую минуту загружается 200 000 фотографий. Каждый месяц выкладывается в открытый доступ 30 млрд новых источников информации.

**Twitter:** Сайт имеет более 645 миллиона пользователей. Каждый день генерируется 175 миллион твитов.

**Google**: Каждую минуту отрабатывается 2,4 миллиона поисковых запросов (40 000 запросов в секунду). Каждый день обрабатывается 25 петабайт данных.

Каждую минуту в мире посылается 204 миллиона е-писем.

По словам специалистов, к категории Big Data относится большинство потоков данных свыше 100 Гб в день.

**Variety (разноообразие).** Возможность воспринимать, хранить и обрабатывать различные данные. Говоря о многообразии, подразумевается следующее.

Различные источники получения данных. Приведем примеры источников возникновения больших данных:

- непрерывно поступающие данные с измерительных устройств,
- события от радиочастотных идентификаторов,
- потоки сообщений из социальных сетей,
- метеорологические данные,
- данные дистанционного зондирования Земли,
- потоки данных о местонахождении абонентов сетей сотовой связи, устройств аудио- и видеорегистрации.

Различные способы представления данных, например, сигналы, поступающие от датчиков, отличаются от текстов научных статей.

Различные форматы хранения (поступления) данных. Это могут быть тексты, аудио- и видео данные, изображения. Более того, одни и те же данные могут быть представлены в различных форматах. Произносимая человеком речь может быть представлена в аудио-формате и в виде текстового файла.

Семантическое разнообразие. Семантика одних и тех же данных может быть представлена по-разному, например, возраст человека может быть указан количественно или в виде таких терминов, как ребенок, юноша, взрослый человек.

Различная степень структурированности данных. Традиционные базы данных позволяют хранить структурированные данные, но фактически в настоящее время порождаемые данные на 80 % являются слабо структурированными или даже неструктурированными.

Технология Big Data позволяет объединять и обрабатывать данные, обладающие приведенному выше многообразию.

Зикопоулоус [4] предложил добавить еще 2 признака – достоверность и ценность (значимость), таким образом получив «5V»:

**Veracity (достоверность).** Свойство, которое характеризует надежность данных. Технология создания и использования традиционных БД предполагает, что в БД поступают тщательно отобранные и проверенные данные. В Big Data дело обстоит иначе. Исходные данные могут быть «сырыми» (неполными, неточными, нечеткими, расплывчатыми, искаженными), то есть поступают без какой-либо предварительной обработки, они могут быть субъективными, случайными и содержать много «шума». Еще один критерий этой характеристики – степень доверия к поступающим данным. Хотя Big Data предоставляют прекрасные возможности для анализа и принятия решений, однако их ценность во многом зависит от качества исходных данных. Технология Big Data учитывает эту характеристику и позволяет надежно работать с такими данными.

**Value (ценность)**. Когда мы говорим о ценности данных, то подразумеваем их значимость с точки зрения прикладных задач. По расчетам IBS, только 1,5 % накопленных массивов данных имеет информационную значимость. Большое количество данных – это хорошо, но если они не представляют никакого интереса, то они бесполезны.

Со временем стали предлагать дополнительные определяющие характеристики Big Data [5–9], которые получили название «7V» и “10V”. Приведем этот дополнительный список.

**Variability (изменчивость).** Под изменчивостью в Big Data подразумевается ситуация, когда постоянно изменяется смысл данных. Например, это имеет место, когда сбор и обработка данных происходит в процессе анализа естественноязыковых тестов и особенно при переводе с одного языка на другой.

**Volatility (волатильность, актуальность).** Характеристика, которая определяет, какой период времени устаревания данных, когда они становятся нерелевантными или бесполезными. Как долго их надо хранить? До эры Big Data данные могли храниться неопределенно долго, использование для этих целей несколько десятков терабайт не было обременительным. Более того, их можно было хранить в действующей базе данных, не вызывая при этом проблем с производительностью. Однако при наличии Big Data, учитывая характеристики объема и скорости, следует тщательно следить за волатильностью данных. Необходимо установить правила управления хранением данных с тем, чтобы обеспечить эффективное их использование.

**Vulnerability (уязвимость).** Большие данные порождают новые проблемы их безопасности. Взлом больших данных приводит к большому взлому. Примером может служить взлом базы данных социальной сети LinkedIn, в результате которого было выкрадено167 млн учетных записей и 360 миллионов сведений о e-mail.

**Validity (пригодность, обоснованность).** Эта характеристика тесно связана с достоверностью и характеризует, в какой мере располагаемые данные являются точными и правильными с точки зрения их предполагаемого использования. По оценке Forbes [10] ученые следующим образом тратят свое время для работы с данными:

– сбор данных 19 %,

– очистка и систематизация данных – 60 %,

– подбор тестовых данных – 3 %,

– анализ данных для построения модели – 9 %,

– уточнение алгоритмов 4 %,

– другие виды работ с данными 5 %.

Таким образом, ученый тратит 80 % своего времени на подбор и подготовку данных прежде, чем приступить к их анализу. Преимуществом использования больших данных для проведения аналитических исследований можно в полной мере воспользоваться только тогда, когда данные тщательно отобраны, являются релевантными и достоверными.

**Visualization (визуализация).** После получения и обработки данных их надо представить таким образом, чтобы они были читабельными и доступными. Именно это и подразумевает визуализация

Как было уже отмечено, в вебе имеет множество определений Big Data. В частности, по адресу [11] дается 43 определения Big Data. Обобщая эти материалы, дадим следующие определение.

**Big Data (большие данные)** – это огромные объемы неоднородной, неструктурированной или слабо структурированной, существенно распределенной и интенсивно растущей, изменяющейся и используемой цифровой информации, которую невозможно обработать традиционными средствами. А также методы, технологии и средства их сбора, хранения и обработки и анализа с целью получения воспринимаемых человеком результатов.

## 4. КлассификацияBig Data

Редактор журнала Web 2.0 Journal Дайон Хинчклифф (Dion Hinchcliffe) дал классификацию Big Data [12], позволяющую соотнести технологию с результатом, который ждут от обработки Big Data. Хинчклиф делит подходы к Big Data на три группы: Fast Data **(**быстрые данные), их объем измеряется терабайтами-петабайтными; Big Analytics **(**большая аналитика) – петабайтные-экзабайтные данные и Deep Insight (глубокое проникновение) – экзабайты-зеттабайты. Группы различаются между собой не только оперируемыми объемами данных, но и качеством решения задач по их обработке.

Обработка для **Fast Data** не предполагает получения новых знаний, ее ре-

зультаты соотносятся с априорными знаниями и позволяют судить о том, как протекают те или иные процессы, она позволяет лучше и детальнее увидеть происходящее, подтвердить или отвергнуть какие-то гипотезы. Только небольшая часть из существующих сейчас технологий подходит для решения задач Fast Data, в этот список попадают некоторые технологии работы с хранилищами. Скорость работы этих технологий должна возрастать синхронно с ростом объемов данных.

Задачи, решаемые средствами **Big Analytics,** заметно отличаются, причем не только количественно, но и качественно, а соответствующие технологии должны помогать в получении новых знаний — они служат для преобразования зафиксированной в данных информации в новое знание. Однако на этом среднем уровне не предполагается наличие искусственного интеллекта при выборе решений или каких-либо автономных действий аналитической системы — она строится по принципу «обучения с учителем». Иначе говоря, весь ее аналитический потенциал закладывается в нее в процессе обучения.

Высший уровень, **Deep Insight**, предполагает обучение без учителя (unsupervised learning) и использование современных методов аналитики, а также различные способы визуализации. На этом уровне возможно обнаружение знаний и закономерностей, априорно неизвестных.

Далее на рис. 1 показана схема Дайон Хинчклиффа взаимодействия трех составляющих Big Data.

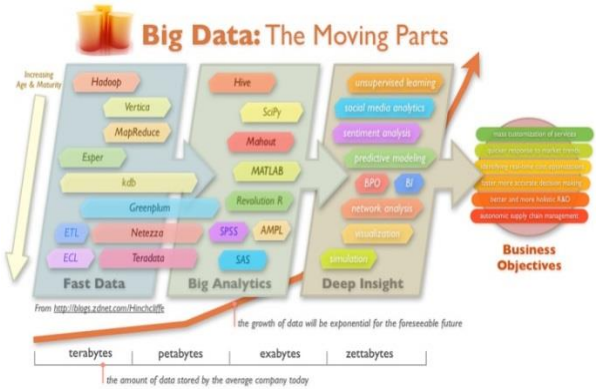


Рис. 1

## 5. Некоторые вехи в истории развития Big Data

Широкое использование термина «большие данные» связывают с Клиффордом Линчем (Clifford Lynch), редактором журнала Nature, подготовившим к 3 сентября 2008 года специальный выпуск номера старейшего британского научного журнала, посвященный поиску ответа на вопрос «Как могут повлиять на будущее науки технологии, открывающие возможности работы с большими объёмами данных?», в котором были собраны материалы о феномене взрывного роста объёмов и многообразия обрабатываемых данных и технологических перспективах в парадигме вероятного скачка «от количества к качеству»; термин был предложен по аналогии с расхожими в деловой англоязычной среде метафорами «большая нефть», «большая руда», отражающими не столько количество чего-то, сколько переход количества в качество. Этот специальный номер подытоживает предшествующие дискуссии о роли данных в науке вообще и в электронной науке (e-science) в частности.

Этот термин был сначала введен в академической среде и прежде всего обсуждалась проблема роста и многообразия научных данных, но начиная с 2009 года термин широко распространился в деловой среде.

В 2010 году появляются первые продукты и технологии, относящиеся исключительно и непосредственно к проблеме обработки больших данных.

К 2011 году большинство крупнейших поставщиков информационных технологий в своих деловых стратегиях начинают использовать понятие «большие данные», это, в частности, относится к IBM, Oracle, Microsoft, Hewlett-Packard, EMC, а основные аналитики рынка информационных технологий посвящают концепции специальные исследования.

Большой шум вокруг темы больших данных возник после того, как в июне 2011 года консалтинговая компания McKinsey выпустила доклад «Большие данные: следующий рубеж в инновациях, конкуренции и производительности», в котором

оценила потенциальный рынок больших данных в миллиарды долларов.

В этом же году аналитическая компания Gartner отметила большие данные как тренд номер два в информационно-технологической инфраструктуре (после виртуализации и как более существенный, чем энергосбережение и мониторинг). В это же время прогнозировалось, что технология больших данных окажет наибольшее влияние на информационные технологии, в производстве, здравоохранении, торговле, государственном управлении.

В 2012 году администрация президента США выделила 200 миллионов долларов для того, чтобы различные американские ведомства организовывали конкурсы по внедрению технологий больших данных в жизнь. Если в 2009 году американские венчурные фонды вложили в отрасль всего 1,1 миллиарда долларов, то в 2012 – уже 4,5 миллиарда долларов.

С 2013 года большие данные как академический предмет начинают изучать в появившихся вузовских программах по науке о данных и вычислительным наукам и инженерии.

В 2015 году Gartner исключил большие данные из цикла зрелости новых технологий и прекратил выпускать выходивший в 2011–2014 годы отдельный цикл зрелости технологий больших данных, мотивировав это переходом от этапа шумихи к практическому применению.

## 6. Принципы работы с Big Data

Исходя из определения Big Data, можно сформулировать следующие основные принципы работы с такими данными [13]:

**распределенность.** Хранить информацию в одном месте бессмысленно и практически невозможно. Поэтому технология работы с Big Data должна использовать распределенное хранение, управление, обработку и анализ данных, хранящихся в разнообразных хранилищах данных во всем мире;

**горизонтальная масштабируемость.** Поскольку данных может быть сколь угодно много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой. В 2 раза вырос объём данных – в 2 раза увеличили кластер и всё продолжило работать с такой же производительностью;

**отказоустойчивость.** Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много. Например, Hadoop-кластер Yahoo имеет более 42000 машин. Это означает, что часть этих машин будет гарантированно выходить из строя. Методы работы с большими данными должны учитывать возможность таких сбоев и переживать их без каких-либо значимых последствий;

**локальность данных**. В больших распределённых системах данные распределены по большому количеству машин. Если данные физически находятся на одном сервере, а обрабатываются на другом – расходы на передачу данных могут превысить расходы на саму обработку. Поэтому одним из важнейших принципов проектирования BigData-решений является принцип локальности данных – по возможности обрабатываем данные на той же машине, на которой они хранятся;

**интерпретация данных в процессе их обработки** (schema-on-read). Данные поступают в хранилище такими, как есть, без какого-либо их предварительного описания, без указания их структуры и семантики. И только в процессе их выборки для обработки происходит их «осмысливание».

Все современные средства работы с большими данными так или иначе следуют этим пятерым принципам.

## 7. Методы и технологии анализа и визуализации, применимые к Big Data

К настоящему времени создано и адаптировано множество методов и технологий для сбора, агрегирования, манипулирования, анализа и визуализации больших данных. Эти методы и технологии заимствованы из различных областей, включая статистику, информатику, прикладную математику и экономику. Это означает, что для извлечения выгоды из

больших данных, следует использовать гибкий междисциплинарный подход. Некоторые методы и технологии были разработаны для оперирования значительно меньшими объемами и разнообразием данных, но были успешно адаптированы для Big Data. Другие были разработаны в последнее время, в частности, для сбора и анализа больших данных. Далее приводится перечень и краткое описание методов и технологий анализа и визуализации, применимые к Big Data , которые взяты из отчета McKinsey [14].

## 7.1. Методы анализа Big Data

**Методы класса Data Mining**:

– **обучение ассоциативным** правилам (association rule learning) – это метод, базирующийся на правилах, используется для обучения машин способам обнаружения зависимостей между данными в больших базах данных;

– **классификация** – методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным;

– **кластерный анализ** – статистический метод классификации объектов, который приводит к разделению разнообразных групп на более мелкие группы подобных (сходных) объектов, для которых критерий подобия заранее не известен;

– **регрессионный анализ**.

**Краудсорсинг** (crowdsourcing) – метод сбора, категоризация и обогащение данных силами широкого круга лиц, привлечённых на основании публичной оферты, без вступления в трудовые отношения, обычно посредством использования сетевых медиа.

**Смешение и интеграция данных** (data fusion and integration) – набор методов, позволяющих интегрировать и анализировать разнородные данные из разнообразных источников для глубинного анализа более точно и эффективно, чем из единственного источника данных. В качестве примеров методов этого класса является цифровая обработка сигналов и обработка естественного языка.

**Обучение ассоциативным правилам** (association rule learning). Совокупность методов для анализа необходимых взаимосвязей, то есть «ассоциативных правил»," среди переменных в больших базах данных.

**Машинное обучение** (machine learning). Класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Включает обучение с учителем (supervised learning) и без учителя (unsupervised learning), а также Ensemble learning – использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей (constituent models).

**Обработка естественного языка** (Natural language processing – NLP). Общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза естественных языков. Применительно к искусственному интеллекту анализ означает понимание языка, а синтез – генерацию грамотного текста. Многие NLP-методы являются методами машинного обучения.

**Искусственные нейронные сети** (artificial neural networks). Математическая модель, построенная по принципу организации и функционирования биологических нейронных сетей – сетей нервных клеток живого организма.

**Сетевой анализ** (network analysis). Набор методов, используемых для описания и анализа отношений между дискретными узлами в графе или сети. В анализе социальной сети анализируются связи между людьми в сообществе или организации, например, как перемещается информация или кто имеет наибольшее влияние на кого.

**Распознавание образов** (pattern recognition). Набор методов машинного обучения, развивающих основы и методы классификации и идентификации предметов, явлений, процессов, сигналов, ситуаций и т. п. объектов, которые характери-

зуются конечным набором некоторых свойств и признаков.

**Прогнозная аналитика** (predictive analytics). Класс методов анализа данных, концентрирующийся на прогнозировании будущего поведения объектов и субъектов с целью принятия оптимальных решений.

**Анализ тональности текста** (sentiment analysis). Класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов (мнений) по отношению к объектам, речь о которых идёт в тексте.

**Имитационное моделирование**. (simulation modeling) – метод исследования, при котором изучаемая система заменяется моделью, с достаточной точностью описывающей реальную систему (построенная модель описывает процессы так, как они проходили бы в действительности), с которой проводятся эксперименты, с целью получения информации об этой системе.

**Пространственный анализ** (Spatial analysis) – набор методов, которые анализируют топологические, геометрические или географические свойства, представленные в наборе данных. Часто данные для пространственного анализа поступают из географических информационных систем (ГИС).

**Статистический анализ**, примеры: A/B-тестирование (контрольная группа элементов сравнивается с набором тестовых групп, в которых один или несколько показателей были изменены, для того, чтобы выяснить, какие из изменений улучшают целевой показатель) и анализ временных рядов.

**Анализ временных рядов** (time series analysis) – совокупность математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и для их прогнозирования. Сюда относятся, в частности, методы регрессионного анализа. Выявление структуры временного ряда необходимо для того, чтобы построить математическую модель того явления, которое является источником анализируемого временного ряда.

## 7.2. Технологии и средства работы с Big Data

Существует множество технологий для агрегации, манипулирования, управления и анализа больших данных. Далее приводится список наиболее известных и используемых технологий и средств. Они приводятся в алфавитном порядке.

Big Table. Запатентованная распределенная система баз данных, построенная на основе Google File System.

**Business intelligence** (BI) (бизнес-аналитика). Совокупность методологий, процессов, архитектур и технологий, которые преобразуют большие объемы «сырых» данных в осмысленную и полезную информацию, пригодную для бизнес-анализа и для поддержки принятия оптимальных тактических и стратегических решений.

**Cassandra.** Свободно распространяемая система управления базами данных, предназначенная для манипулирования данными огромного объема в распределенных системах.

**Cloud computing** (облачные вычисления). Вычислительная парадигма, в которой высокомасштабируемые вычислительные ресурсы, обычно сконфигурированные в виде распределенных систем, предоставляются в сетях качестве сервисов.

**Data Warehouse** (хранилище данных). Предметно-ориентированная информационная база данных, специально разработанная и предназначенная для подготовки отчётов и анализа данных с целью поддержки принятия решений в организации и является одной из основных компонент бизнес-анализа. Выступает центральным репозиторием данных, поступающих из различных источников. Хранит текущие и исторические данные. Строится на базе систем управления базами данных и систем поддержки принятия решений.

**Distributed system** (распределенная система). Множество компьютеров, взаи-

модействующих по сети и объединенных для решения общей вычислительной задачи.

**Dynamo.** Запатентованная распределенная система хранения данных, разработанная в Amazon.

**Extract, transform, and load** (ETL) (извлечь, преобразовать, загрузить). ПОР, используемое для извлечения данных из внешних источников, преобразования их для удовлетворения операционных потребностей, и загрузка их в базу данных или хранилище данных.

**Google File System.** Запатентованная распределенная файловая система. На ее основе построен Hadoop.

**Hadoop**. Проект фонда Apache Software Foundation, свободно распространяемый набор утилит, библиотек ифреймворк для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов. Используется для реализации поисковых и контекстных механизмов многих высоконагруженных веб-сайтов, в том числе, для Yahoo! и Facebook. Базируется на MapReduce и Google File System.

HBase. Свободно распространяемая распределенная нереляционная база данных, созданныя на основе Big Table Google.

**MapReduce**. Модель распределённых вычислений, представленная компанией Google, используемая для параллельных вычислений над очень большими, вплоть до нескольких петабайт, наборами данных в компьютерных кластерах. Эта моделоь реализована в Hadoop.

**Mashup**. Веб-приложение, объединяющее данные из нескольких источников в один интегрированный, например, при объединении картографических данных Google Maps с данными о недвижимости с Craigslist получается новый уникальный веб-сервис, изначально не предлагаемый ни одним из источников данных.

R. Свободно распространяемый язык программирования среда программирования для статистических и графических вычислений.

Stream processing. Технология, предназначенная для обработки больших потоков данных в реальном масштабе времени.

## 7.3. Визуализация Big Data

Наглядное представление результатов анализа больших данных таким образом, чтобы ее можно было легко воспринимать, является ключевой проблемой анализа данных, имеет принципиальное значение для их интерпретации. Восприятие человека ограничено, и ученые продолжают вести исследования в области совершенствования современных методов представления данных в виде изображений, диаграмм или анимаций. В качестве иллюстрации приводим несколько прогрессивных методов визуализации, относительно недавно получивших распространение.

**Облако тегов** (Tag cloud) рис. 2. Каждому элементу в облаке тегов присваивается определенный весовой коэффициент, который коррелирует с размером шрифта. В случае анализа текста величина весового коэффициента напрямую зависит от частоты употребления (цитирования) определенного слова или словосочетания. Позволяет читателю в сжатые сроки получить представление о ключевых моментах сколько угодно большого текста или набора текстов.


Рис. 2

**Clustergram** (кластерграмма) рис. 3. Метод визуализации, использующийся при кластерном анализе. Показывает, как отдельные элементы множества данных

соотносятся с кластерами по мере изменения их количества. Выбор оптимального количества кластеров – важная составляющая кластерного анализа. Этот способ визуализации позволяет аналитику лучше понять, как результаты кластеризации изменяются по мере изменения количества кластеров.

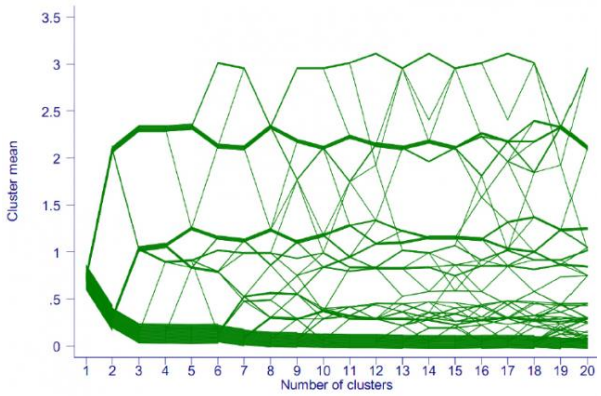


Рис. 3

**History flow** (исторический поток) рис. 4. Помогает следить за эволюцией документа, над созданием которого работает одновременно большое количество авторов. В History flow (исторический поток). Помогает следить за эволюцией документа, над созданием которого работает одновременно большое количество авторов. В частности, это типичная ситуация для сервисов wiki. По горизонтальной оси откладывается время, по вертикальной – вклад каждого из соавторов, т. е. объем введенного текста. Каждому уникальному автору присваивается определенный цвет на диаграмме. Приведенная диаграмма – результат анализа для слова «ислам» в Википедии. Хорошо видно, как возрастала активность авторов с течением времени.

**Spatial information flow** (пространственный поток) рис. 5. Эта диаграмма позволяет отслеживать пространственное распределение информации. Приведенная в качестве примера диаграмма построена с помощью сервиса New York Talk Exchange. Она визуализирует интенсивность обмена IP-трафиком между Нью-Йорком и другими городами мира. Чем ярче линия – тем больше данных передается за единицу времени. Таким образом, не составляет труда выделить регионы, наиболее близкие к Нью-Йорку в контексте информационного обмена.

Рис. 4

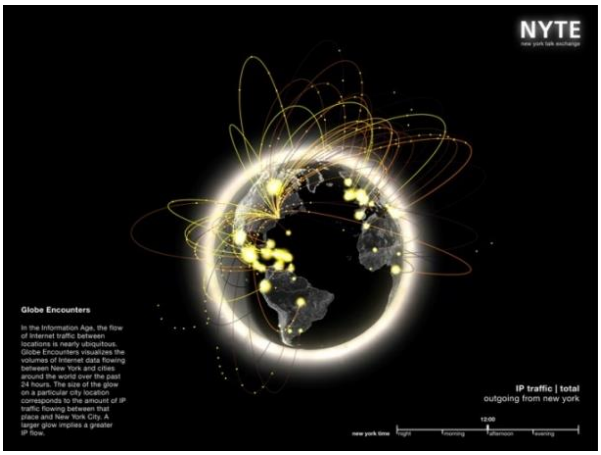


Рис. 5

## 8. Жизненный цикл управления данными с использованием технологии Big Data

Опишем в общих чертах жизненный цикл управления данных, который использует технологию Big Data. Идея этого цикла взята из работы [15] Предлагаемый жизненный цикл данных состоит из следующих этапов: сбор, фильтрация и классификация, анализ данных, хранение, обмен и публикация, а также поиск и обнаружение данных. Далее кратко описывается каждый этап согласно показанному на рис. 6. жизненному циклу.

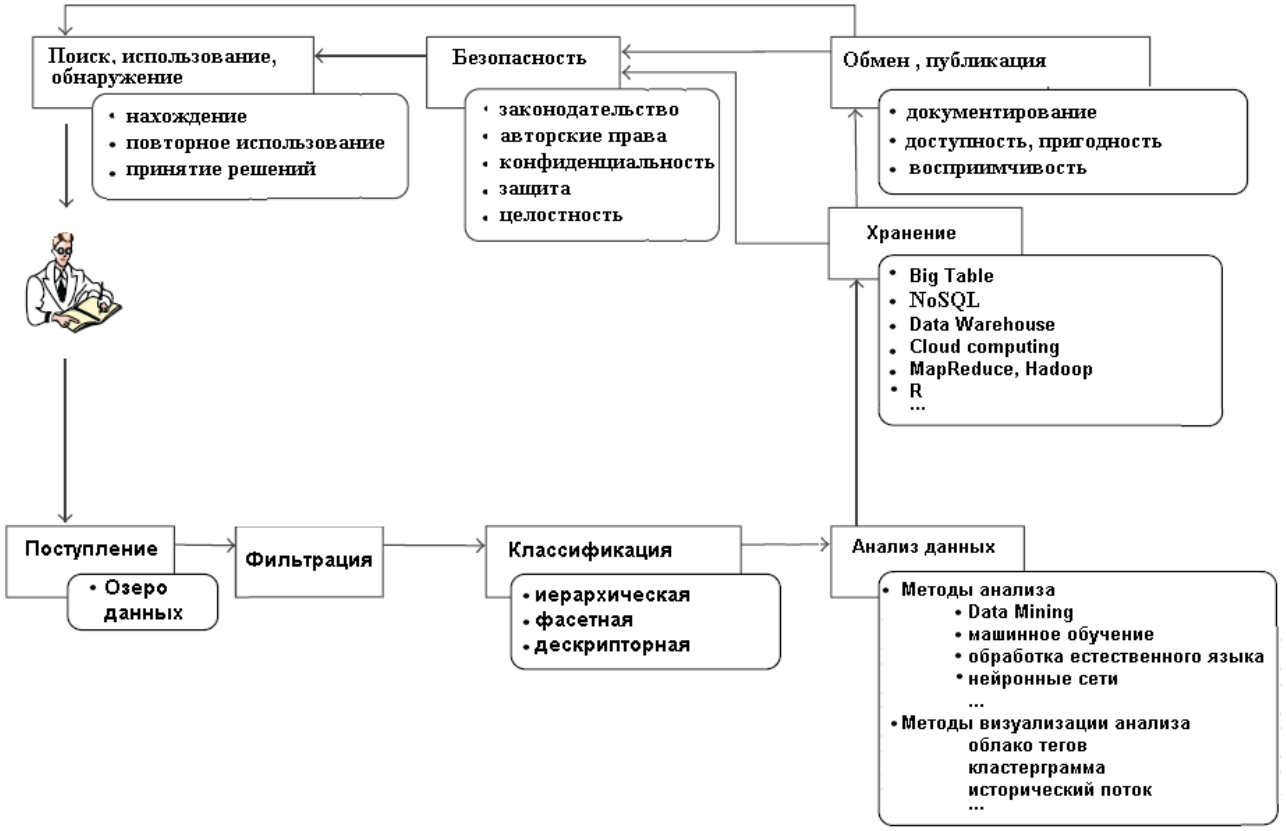


Рис. 6

**8.1. Поступление данных**. Поступление (сбор) данных – это первый этап жизненного цикла данных. Большое количество данных поступает из различных источников. Такими источниками могут быть: файлы журналов, которые ведутся на серверах, датчики различного вида, мобильные устройства, данные, поступающие со спутников, результаты научных исследований, данные вычислительных экспериментов, результаты выполнения поисковых запросов, данные, порождаемые в социальных сетях, и многие другие. При сборе данных используются разнообразные методы получения исходных сырых данных из различных источников. Рассмотрим несколько методов сбора данных и используемые ими технологии.

– Файлы журналов (log-файлы). Этот метод используется для автоматической регистрации данных, связанных с различными событиями, происходящими в автоматизированных системах. Log-файлы используются практически во всех компьютерных системах, например, веб-сервера фиксирует все транзакции, выполняемые сервером. При наличии очень больших файлов журнала их информация запоминается в базах данных, а не в виде тестовых файлов.

– Сенсорные данные (Sensor data.). Часто датчики используются для съема физических характеристик, которые затем преобразуются в воспринимаемые цифровые сигналы для их сохранения и обработки. К сенсорным данным можно отнести, например, данные, которые поступают в виде звуковых, вибрационных, голосовых волн, результатов физических, химических, биологических, метеорологических или других видов исследований, результатов съема характеристик (показателей) производственных процессов.

– Мобильные устройства. С помощью различные технологий, которые встраиваются в мобильные устройства, можно получать и передавать информацию географическом местоположении, воспринимать аудио- и видеоинформацию, делать фотографии, с помощью сенсорных экранов и гравитационных датчиков получать

информацию о состоянии здоровья человека.

В результате сбора таких данных образуется так называемое озеро данных (Data lake). Это централизованное хранилище больших данных в сыром, необработанном виде. В нем хранят данные из разных источников, разных форматов, структурированные, слабо структурированные, неструктурированные и бинарные данные (изображения, аудио видео-данные)). Они хранятся как правило, в несистематизированном виде такими, как есть, без какой либо предварительной обработки. Это обходится значительно дешевле традиционных хранилищ, в которые помещаются только структурированные данные. Data lake позволяют анализировать большие данные в исходном виде.

**8.2. Фильтрация данных.** В исходных данных может быть много шума. Так, например, при некачественной аудиозаписи фоновый шум может быть настолько сильным, что не позволяет выделить полезную аудио-информацию с использование современных средств распознавания, или камера видео-наблюдения произвела съемку в темное время и изображение абсолютно черным. Фильтрация позволяет ибавиться от такой информации.

**8.3. Классификация данных.** Любые поступающие данные всегда обладают какой-то минимальной информацией. Например, известно, где именно установлена видео-камера, куда она направлена и к какому времени суток привязаны те или иные кадры, или что собой представляют поступающие научные данные, результатами какого эксперимента они являются, при каких условиях эксперимент проводился, и так далее. Таким образом, любые поступающие данные обладают так называемыми метаданными, которые можно использовать для проведения первоначальной классификации, которая является первоначальным шагом выявления семантики данных. Эта семантика служит хорошей основой для проведения последующего анализа данных.

Методы классификации – это совокупность приемов разделения множества объектов на подмножества. В науке известны три метода классификации объектов: иерархический, фасетный, дескрипторный. Эти методы различаются разной стратегией применения классификационных признаков.

**Иерархический метод.** Это метод, при котором заданное множество последовательно делится на подчиненные подмножества, постепенно конкретизируя объект классификации. При этом основанием деления служит некоторый выбранный признак. Совокупность получившихся группировок при этом образует иерархическую древовидную структуру.

**Фасетный метод**. Подразумевает параллельное разделение множества объектов на независимые классификационные группы. При этом не предполагается жёсткой классификационной структуры и заранее построенных конечных групп. Классификационные группировки образуются путём комбинации значений, взятых из соответствующих фасетов.

**Дескрипторный метод.** Суть этого метода заключается в следующем: отбирается совокупность ключевых слов или словосочетаний, описывающих определенную предметную область или совокупность однородных объектов, они подвергаются нормализации, на основании этого создается словарь дескрипторов, который служит основой для проведения классификации.

**8.4. Анализ данных.** Анализ данных позволяет воспринять и обработать огромные объемы Big Data. Анализ данных является сложной задачей и во многом зависит от тех задач, которые надо решать с использованием этих данных, выдвигаемых требований к точности и скорости решения, наличия технических средств и, наконец, состояний исходных данных. Анализ данных включает решения следующих двух основных задач:

– на первом этапе должна быть решена задача раскрытия синтаксиса данных, то есть выявление структуры данных, например, какие объекты предоставляемые данные представляют, какими свойствами они обладают, что собой

представляют значения этих свойств, каким образом взаимосвязаны объекты, какова природа и каковы характеристики этих связей;

– второй этап связан с раскрытием семантики данных. Это так называемый этап интеллектуального анализа данных (data mining). В разделе «Методы анализа Big Data» приводится краткое описание используемых методов. Для гибкой организации анализа данных в работе [16] были предложены следующие три принципа: во-первых, для достижения поставленных целей следует использовать не единственный, а множество релевантных методов анализа. Во-вторых, для хранения данных следует использовать различные методы и устройства хранения, которые могут быть распределены по компьютерам сети. В-третьих, следует предоставлять высокоэффективные методы и средства доступа и обработки данных.

Анализ данных производится с учетом следующих факторов: гетерогенность, точность и сложность данных, возможность их масштабирования.

**8.5. Хранение, совместное использование, публикация**. После сбора, очистки и анализа полученные данные запоминаются в соответствующих хранилищах, к ним предоставляется доступ и/или они публикуются для ознакомления с ними широкого круга заинтересованных лиц. Большие по объему и интенсивно использующиеся наборы данных. Big Data должны храниться и управляться с большой степенью надежности, доступности и простоте использования. Инфраструктура хранения должна обладать достаточной степенью гибкости. Система хранения должна быть распределенной. Такая распределенная система хранения должно обеспечить поддержку целостности, обеспечение доступности, устойчивости к отказам различного вида.

**8.6. Безопасность.** Безопасность данных – это защита данных от несанкционированного (случайного или намеренного) доступа, изменения или разрушения.

Сфера применения Big Data в современном мире практически не имеет границ. Раскрытие, изменение или разрушение данных в Big Data может иметь катастрофические последствия. При этом следует отметить, что все среды для работы с большими данными подвержены рискам В связи с этим необходимо обеспечивать надежную защиту Big Data при их хранении, передаче и обработке за счет внедрения и использования процедур и технологических решений в области защиты информации.

**8.7. Поиск, повторное использование, обнаружение.** Поиск данных обеспечивает (гарантирует) качество данных, увеличение их значимости и сохранности посредством механизма повторного использования и сохранения с целью выявления новой более осмысленной информации. Сфера этой деятельности включает поиск, обнаружение, управление, аутентификацию, архивирование, сохранение и представление данных .После публикации данных другие исследователи должны иметь возможность аутентифицировать и регенерировать их в соответствии со своими интересами для проведения своих исследований. Возможность повторного использования опубликованных данных также должна быть гарантирована в научных сообществах. При многократном использовании определение семантики опубликованных данных является обычной ситуацией. Обычно эта процедура выполняется вручную. В Европейском Союзе активно поддерживается концепция открытой науки, например, инициированием Европейского облака открытой науки для обеспечения открытого доступа к результатам научных исследований из финансируемых государством проектов.

## Литература

1. Chernyuk L. Big Data – new theory and practice. Otkrytye sistemy. SUBD 2011 № 10. URL: https://www.osp.ru/os/2011/10/13010990/

2. Laney Doug (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety. Technical Report 949, METAGroup (now Gartner). [Electronic resourse]: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
3. Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani1. Big Data: Survey, Technologies, Opportunities, and Challenges // Hindawi Publishing Corporation The Scientific World Journal Volume 2014, Article ID 712826, 18 pages, URL: http://dx.doi.org/10.1155/2014/712826
4. Zikopoulos P., Parasuraman K., Deutsch T., Giles J., Corrigan D. (2013) Harness the power of big data The IBM big data platform. McGraw Hill Professional, New York, NY. - [Электронный ресурс]: ftp://public.dhe.ibm.com/software/pdf/at/SWP10/Harness_the_Power_of_Big_Data.pdf
5. The Four V′s of Big Data (англ.). IBM (2011). Проверено 19 февраля 2017. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg
6. Neil Biehn. The Missing V's in Big Data: Viability and Value (англ.). Wired (1 May 2013). Проверено 19 февраля 2017. https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/
7. Eileen McNulty. Understanding Big Data: The Seven V's (англ.). Dataconomy (22 May 2014). Проверено 19 февраля 2017. http://dataconomy.com/2014/05/seven-vs-big-data/
8. Tom McNeill. The Eight V's of Supercomputing and Big Data. https://www.nimbix.net/eight-vs-supercomputing-big-data/
9. George Firican. The 10 Vs of Big Data - https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
10. Gil Press. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/
11. Jennifer Dutcher. What Is Big Data? https://datascience.berkeley.edu/what-is-big-data/
12. Dion Hinchcliffe. Big Data, The Moving Parts: Fast Data, Big Analytics, and Deep Insight. - https://www.flickr.com/photos/dionh/7550578346/in/photostream/
13. Big Data от А до Я. Часть 1: Принципы работы с большими данными, парадигма MapReduce. https://habr.com/company/dca/blog/267361/
14. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf
15. Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani. Big Data: Survey, Technologies, Opportunities, and Challenges. Hindawi Publishing Corporation The Scientific World Journal, Volume 2014, Article ID 712826, 18 pages. URL: http://dx.doi.org/10.1155/2014/712826
16. E. Begoli and J. Horey, "Design principles for effective knowledge discovery frombig data," in Proceedings of the 10thWorking IEEE/IFIP Conference on Software Architecture (ECSA '12). P. 215–218, August 2012.

## Referenses

1. Chernyuk L. Big Data – new theory and practice. Otkrytye sistemy. SUBD 2011 № 10. URL: https://www.osp.ru/os/2011/10/13010990/
2. Laney Doug (2001) 3D Data Management: Controlling Data Volume, Velocity, and Variety. Technical Report 949, METAGroup (now Gartner). [Electronic resourse]: https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf
3. Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani1. Big Data: Survey, Technologies, Opportunities, and Challenges // Hindawi Publishing Corporation The Scientific World Journal Volume 2014, Article ID 712826, 18 pages, URL: http://dx.doi.org/10.1155/2014/712826
4. Zikopoulos P., Parasuraman K., Deutsch T., Giles J., Corrigan D. (2013) Harness the power of big data The IBM big data platform.

McGraw Hill Professional, New York, NY. - [Электронный ресурс]: ftp://public.dhe.ibm.com/software/pdf/at/SWP10/Harness_the_Power_of_Big_Data.pdf
5. The Four V´s of Big Data (англ.). IBM (2011). Проверено 19 февраля 2017. http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg
6. Neil Biehn. The Missing V's in Big Data: Viability and Value (англ.). Wired (1 May 2013). Проверено 19 февраля 2017. https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value/
7. Eileen McNulty. Understanding Big Data: The Seven V's (англ.). Dataconomy (22 May 2014). Проверено 19 февраля 2017. http://dataconomy.com/2014/05/seven-vs-big-data/
8. Tom McNeill. The Eight V's of Supercomputing and Big Data. https://www.nimbix.net/eight-vs-supercomputing-big-data/
9. George Firican. The 10 Vs of Big Data - https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
10. Gil Press. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/
11. Jennifer Dutcher. What Is Big Data? https://datascience.berkeley.edu/what-is-big-data/
12. Dion Hinchcliffe. Big Data, The Moving Parts: Fast Data, Big Analytics, and Deep Insight. - https://www.flickr.com/photos/dionh/7550578346/in/photostream/
13. Big Data from A to Z. Part 1: Principles of working with Big Data, paradigm MapReduce. https://habr.com/company/dca/blog/267361/
14. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, Angela Hung Byers. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011. https://bigdatawg.nist.gov/pdf/MGI_big_data_full_report.pdf
15. Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker Targio Hashem, Zakira Inayat, Waleed KamaleldinMahmoud Ali, Muhammad Alam, Muhammad Shiraz, and Abdullah Gani. Big Data: Survey, Technologies, Opportunities, and Challenges. Hindawi Publishing Corporation The Scientific World Journal, Volume 2014, Article ID 712826, 18 pages. URL: http://dx.doi.org/10.1155/2014/712826
16. E. Begoli and J. Horey, "Design principles for effective knowledge discovery frombig data," in Proceedings of the 10thWorking IEEE/IFIP Conference on Software Architecture (ECSA '12). P. 215–218, August 2012.

Получено 05.07.2019

***Об авторе:***

*Резниченко Валерий Анатольевич,*
кандидат физико-математических наук,
старший научный сотрудник Института программных систем НАН Украины.
Количество научных публикаций в украинских изданиях – 61.
Количество научных публикаций в зарубежных изданиях – 4.
http://orcid.org/0000-0002-4451-8931

***Место работы автора***:

Институт программных систем
НАН Украины.
03187, Киев,
проспект Академика Глушкова, 40.
Тел.: +38 (044) 526 5139.
E-mail: vreznichenko_47@mail.ru