ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р ИСО/МЭК 20546-2019

Информационные технологии

БОЛЬШИЕ ДАННЫЕ

Обзор и словарь

(ISO/IEC 20546:2019, IDT)

Издание официальное

Настоящий проект стандарта не подлежит применению до его утверждения

Москва
Стандартинформ
2020

Предисловие

1 РАЗРАБОТАН Национальным центром цифровой экономики МГУ имени М.В. Ломоносова и автономной некоммерческой организацией «Институт развития информационного общества».

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164«Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от №

4 Настоящий стандарт идентичен международному стандарту ISO (ИСО) 20546:2019 «Информационные технологии – Большие данные – Обзор и словарь» (ISO/IEC 20546:2019 "Information technology — Big data — Overview and vocabulary", IDT)

5 ВВЕДЕН ВПЕРВЫЕ

Правила применения настоящего стандарта установлены в статье 26 Федерального закона от 29 июня 2015 г. № 162-ФЗ «О стандартизации в Российской Федерации». Информация об изменениях к настоящему стандарту публикуется в ежегодном (по состоянию на 1 января текущего года) информационном указателе «Национальные стандарты», а официальный текст изменений и поправок — в ежемесячном информационном указателе «Национальные стандарты». В случае пересмотра (замены) или отмены настоящего стандарта соответствующее уведомление будет опубликовано в ближайшем выпуске ежемесячного информационного указателя «Национальные стандарты». Соответствующая информация, уведомление и тексты размещаются также в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.gost.ru)

Содержание

Введение

Установленные в настоящем стандарте термины и определения расположены в систематизированном порядке, отражающем систему понятий данной области знания.

Для каждого понятия установлен один стандартизованный термин.

В стандарте приводятся иноязычные эквиваленты стандартизованных терминов на английском (en) языке.

В стандарте приведен алфавитный указатель терминов на русском языке

Стандартизованные термины набраны полужирным шрифтом, их краткие формы - светлым, а недопустимые термины-синонимы - курсивом.

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Информационные технологии

БОЛЬШИЕ ДАННЫЕ

Обзор и словарь

Information technology – Big data reference architecture – Overview and vocabulary

Дата введения – 2020

1 Область применения

В настоящем стандарте устанавливаются термины и определения (буквенные обозначения) понятий в области информационных технологий и больших данных.

Данный документ содержит набор терминов и определений, необходимых для улучшения информационного взаимодействия и понимания информационных технологий и больших данных. Он обеспечивает терминологическую основу для стандартов, связанных с большими данными.

Термины, установленные настоящим стандартом, обязательные для применения во всех видах документации и литературы по данной научно-технической отрасли, входящей в сферу работ по стандартизации и (или) использующих результаты этих работ.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ISO/IEC 2382:2015, Information technology — Vocabulary (Информационные технологии. Словарь) ГОСТ 33707-2016 (ISO/IEC 2382:2015) Информационные технологии (ИТ). Словарь

ISO 9075 (all parts), Information technology — Database languages — SQL (Информационные технологии. Язык базы данных. Язык структурированных запросов)

ISO/IEC 11404, Information technology — General-Purpose Datatypes (GPD) (Информационные технологии. Типы данных общего назначения)

ISO/IEC 17788:2014, Information technology — Cloud computing — Overview and vocabulary (Информационные технологии. Облачные вычисления. Обзор и словарь. ГОСТ ISO/IEC 17788-2016 Информационные технологии. Облачные вычисления. Общие положения и терминология)

ISO/IEC 19784-4:2011, Information technology — Biometric application programming interface — Part 4: Biometric sensor function provider interface (Информационные технологии. Биометрический программный интерфейс. Часть 4. Интерфейс поставщика функции биометрического датчика)

ГОСТ 33707-2016 (ISO/IEC 2382:2015) Информационные технологии (ИТ). Словарь

ГОСТ ISO/IEC 17788-2016 Информационные технологии (ИТ). Облачные вычисления. Общие положения и терминология

ГОСТ Р ИСО/МЭК 19784-4-2014 Информационные технологии (ИТ). Биометрия. Биометрический программный интерфейс. Часть 4. Интерфейс поставщика функции биометрического датчика

Примечание - При пользовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования - на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю "Национальные стандарты", который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя "Национальные стандарты" за текущий год. Если ссылочный стандарт заменен (изменен), то при пользовании настоящим стандартом следует руководствоваться заменяющим (измененным) стандартом. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, применяется в части, не затрагивающей эту ссылку.

3 Стандартизованные термины с определениями

Для целей этого документа используются следующие термины и определения.

ISO (ИСО) и IEC (МЭК) поддерживают терминологические базы данных для использования в стандартизации по следующим адресам:

— Онлайн-библиотека стандартов ISO (ИСО): доступна по адресу https://www.iso.org/obp.

— Международный электротехнический словарь МЭК (IEC Electropedia): доступен по адресу http://www.electropedia.org/.

3.1 Термины

3.1.1 **выгода** (benefit)**:** польза для организации от структуризации практических знаний, полученных из аналитической системы.

Примечание - Большие данные часто ассоциируются с выгодой вследствие понимания того, что данные имеют потенциальную ценность, ранее обычно не рассматриваемую.

3.1.2 **большие данные** (big data): большие массивы данных (3.1.11), – главным образом, по таким характеристикам данных (3.1.5), как объем, разнообразие, скорость обработки и/или вариативность, – которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа.

Примечание - Большие данные повсеместно используются множеством различных способов, например, в качестве названия технологии масштабирования, используемой для обработки обширных массивов данных.

3.1.3 **облачные вычисления** (cloud computing): парадигма для обеспечения сетевого доступа к масштабируемому и гибкому пулу совместно используемых физических или виртуальных ресурсов с системой самообслуживания и администрированием по требованию.

Примечание - Примерами таких ресурсов являются серверы, операционные системы, сети, программное обеспечение, приложения и оборудование для хранения.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 17788:2014, 3.2.5]

3.1.4 **кластер** (cluster)**:** (распределенная обработка данных) набор функциональных блоков под общим контролем

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 2120586]

3.1.5 **данные** (data): реинтерпретируемое представление информации в формализованном виде, пригодном для коммуникации, интерпретации или обработки.

Примечание - к записи: Данные могут быть обработаны людьми или автоматическими средствами.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 2121272]

3.1.6 **аналитика данных** (data analytics)**:** составное понятие, состоящее из получения, сбора, проверки и обработки данных (3.1.9), включая их количественную оценку, визуализацию и интерпретацию.

Примечание – Аналитика данных используется для понимания объектов, представленных данными (3.1.5), для прогнозирования конкретных ситуаций и для рекомендаций по шагам для достижения целей. Выводы, полученные из аналитики, используются для различных задач, таких как принятие решений, исследования, устойчивое развитие, проектирование, планирование и т. д.

3.1.7 **база данных** (database): совокупность данных (3.1.5), организованная в соответствии с концептуальной структурой, которая описывает характеристики этих данных и взаимосвязи между их соответствующими объектами, обеспечивая одну или несколько областей применения.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 2121413]

3.1.8 **модель данных** (data model): схема структурирования данных (3.1.5) в базе данных (3.1.7) в соответствии с формальными описаниями в ее информационной системе и требованиями используемой системы управления базой данных.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 2125519]

3.1.9 **обработка данных** (data processing): систематическое выполнение операций с данными (3.1.5).

Примечания

1 Пример: Арифметические или логические операции с данными, объединение или сортировка данных или такие операции с текстом, как редактирование, сортировка, объединение, хранение, извлечение, отображение или печать.

2 Примечание к записи: Термин «обработка данных» не должен использоваться в качестве синонима для обработки информации.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 01.01.06].

3.1.10 **наука о данных** (data science): извлечение практических знаний из данных (3.1.5) посредством исследования или создания и проверки гипотез.

3.1.11 **массив данных** (data set, dataset): идентифицируемая совокупность данных (3.1.5), к которой можно получить доступ или скачать в одном или нескольких форматах.

[ИСТОЧНИК: Адаптировано из Международного стандарта ISO (ИСО) 19115-2:2009, 4.7].

3.1.12 **тип данных** (data type, datatype): определенный массив объектов данных (3.1.5) конкретной структуры данных и набор допустимых операций, в рамках которых эти объекты данных выступают в роли операндов при выполнении любой из этих операций.

Примечания

1 Пример: Целочисленный тип данных имеет очень простую структуру, каждый экземпляр которой, обычно называемый значением, представляет собой член заданного диапазона целых чисел, а допустимые действия включают в себя обычные арифметические операции над этими целыми числами.

2 При отсутствии вероятности двусмысленного толкования вместо термина «тип данных» может использоваться термин «тип».

3 Тип данных: определение и термины, стандартизированные ISO/IEC (ИСО/МЭК) [Международный стандарт ISO/IEC (ИСО/МЭК) 2382-15:1999].

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 2122374]

3.1.13 **вариативность данных** (data variability): изменения в скорости передачи, формате или структуре, семантике или качестве массива данных (3.1.11).

3.1.14 **разнообразие данных** (data variety): диапазон форматов, логических моделей, временных шкал и семантики массива данных (3.1.11).

Примечание - Разнообразие данных относится к нерегулярным или неоднородным структурам данных, их навигации, запросам и типизации данных.

3.1.15 **скорость обработки данных** (data velocity): скорость потока, с которой данные (3.1.5) создаются, передаются, хранятся, анализируются или визуализируются.

3.1.16 **достоверность данных** (data veracity): полнота и/или точность данных (3.1.5).

Примечание – Под достоверностью данных понимаются пояснительные данные и самоанализ объектов для поддержки принятия решений в режиме реального времени.

3.1.17 **изменчивость данных** (data volatility): характеристика данных (3.1.5), относящаяся к скорости изменения этих данных с течением времени.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 17.06.06]

3.1.18 **объем данных** (data volume): степень количества данных (3.1.5), оказывающая влияние на ресурсы для вычислений и хранения, а также на управление ими в процессе обработки данных.

Примечание – Объем данных становится важным при работе с большими массивами данных (3.1.11), включая их.

3.1.19 **распределенная обработка данных** (distributed data processing): обработка данных (3.1.9), в которой выполнение операций распределено между узлами компьютерной сети.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 18.01.08]

3.1.20 **распределенная файловая система** (distributed file system): система, управляющая файлами и папками в нескольких сетевых системах.

3.1.21 **файл** (file): именованный набор записей, рассматриваемый как единое целое.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 04.07.10]

3.1.22 **сборка** (gather): объединение результатов из нескольких узлов в кластере

Примечание к записи: См. распределение/сборка (3.2.33).

3.1.23 **горизонтальное масштабирование** (horizontal scaling): формирование единого логического блока путем соединения нескольких аппаратных и программных средств.

Примечания

1 Примером горизонтального масштабирования является повышение производительности распределенной обработки данных путем добавления узлов в кластере для дополнительных ресурсов.

2 Горизонтальное масштабирование для увеличения производительности также называется масштабированием вширь (scale-out).

3.1.24 **метаданные** (metadata): данные (3.1.5) о данных или элементах данных, которые могут включать в себя их описания, а также данные о владении данными, путях и правах доступа и об изменчивости данных (3.1.17).

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 17.06.05]

3.1.25 **нереляционная база данных** (non-relational database): база данных (3.1.7), не следующая реляционной модели (3.1.31).

Примечание -- «NoSQL», что обычно переводится как «не SQL» или «не только SQL», является общеупотребительным термином для обозначения баз данных, не соответствующих реляционной модели.

3.1.26 **нереляционная модель данных** (non-relational model): логическая модель данных (3.1.10), не следующая реляционной модели (3.1.31) хранения и обработки данных (3.1.5).

3.1.27 **параллельная работа** (parallel): относится к процессу, в котором все события происходят в одном и том же интервале времени, и при этом каждое из них обрабатывается отдельной, но схожей функциональной единицей.

Примечание -- Пример: Параллельная передача битов компьютерного слова по линиям внутренней шины.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 03.02.01]

3.1.28 **частично структурированные данные** (partially structured data): данные (3.1.5), имеющие некую структуру.

Примечания

1. Частично структурированные данные в индустрии часто называют полуструктурированными.

2. Примерами частично структурированных данных являются записи со свободными текстовыми полями в дополнение к более структурированным полям. Такие данные часто представлены в компьютерно-интерпретируемых/разбираемых форматах, таких как XML или JSON.

3.1.29 **реляционная алгебра** (relational algebra): алгебра для выражения и манипулирования отношениями.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 17.04.08]

3.1.30 **реляционная база данных** (relational database): база данных (3.1.7), данные в которой организованы по реляционной модели (3.1.31).

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 17.04.05]

3.1.31 **реляционная модель данных** (relational model): модель данных (3.1.10), структура которой основана на реляционных отношениях.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 17.04.04

3.1.32 **распределение** (scatter): Распределение обработки по нескольким узлам в кластере (3.1.4).

Примечание – См. распределение-сборка (3.2.33).

3.2.33 **распределение-сборка** (scatter-gather): вид обработки больших массивов данных (3.1.11), где необходимые вычисления разделяются и распределяются по нескольким узлам в кластере, а общий результат формируется путем объединения результатов от каждого узла.

Примечание – Обработка методом распределения-сборки обычно требует алгоритмического изменения обрабатывающего программного обеспечения. Примером обработки данных методом распределения-сборки является MapReduce.

3.1.34 **потоковые данные** (streaming data): данные (3.1.5), передаваемые через интерфейс от непрерывно работающего источника.

[ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 19784-4:2011, 4.4]

3.1.35 **структурированные данные** (structured data): данные (3.1.5), организованные на основе предопределенного (применимого) набора.

Примечания

1 Предопределенный набор правил, регулирующих основу для структурирования данных, должен быть четко изложен и опубликован.

2 Предопределенная модель данных часто используется для управления структурированием данных.

3.1.36 **SQL**: язык баз данных, описанный в Международном стандарте ISO/IEC (ИСО/МЭК) 9075.

Примечание – SQL иногда интерпретируется как язык структурированных запросов (Structured Query Language), но это название не используется в серии ISO/IEC (ИСО/МЭК) 9075.

3.1.37 **неструктурированные данные** (unstructured data): данные (3.1.5), характеризуемые отсутствием какой-либо структуры, кроме структуры на уровне записи или файла.

Примечания

1 В целом, неструктурированные данные не состоят из элементов данных.

2 Пример: Примером неструктурированных данных является свободный текст.

3.1.38 **вертикальное масштабирование** (vertical scaling): повышение производительности обработки данных за счет улучшения процессоров, памяти, хранилища или связи.

Примечание – Вертикальное масштабирование для увеличения производительности также называется масштабированием ввысь (scale-up).

3.2 Сокращения

JSON: – Javascript Object Notation (обозначение объектов Javascript)

PII: – Personally Identifiable Information (личная информация)

XML: – Extensible Markup Language (расширяемый язык разметки)

3.3 Ключевые характеристики больших данных

3.3.1 Общие сведения

При выборе системы больших данных необходимо руководствоваться четырьмя характеристиками – объемом, скоростью обработки, разнообразием и вариативностью данных см. 3.3.2. Управление этими характеристиками данных определяется характеристиками обработки в соответствии с описанием в разделе 3.3.3.

3.3.2 Ключевые характеристики данных

3.3.2.1 **Объем данных.** Объем данных представляет собой значительное количество данных, доступных для анализа с целью извлечения полезной информации. Одним из основных факторов развития технологий обработки больших данных стало огромное количество данных, генерируемых с использованием интернета.

3.3.2.2 **Скорость обработки данных.** Скорость обработки данных – это скорость потока создания, хранения, анализа или визуализации данных. Скорость обработки больших данных означает необходимость обработки большого количества данных за короткий промежуток времени. В качестве примера работы с данными с высокой скоростью обработки обычно приводят технологии потоковых данных.

3.3.2.3 **Разнообразие данных.** Разнообразие данных представляет собой необходимость анализа данных разного типа из различных предметных областей. Разнообразные данные обрабатывались путем преобразований или предварительной аналитики для выявления свойств, позволяющих интеграцию их с другими данными. Более широкий диапазон форматов данных, логических моделей, временных шкал и семантики, которые желательно использовать при аналитике данных, усложняет интеграцию разнообразных данных. В качестве средства, способствующего интеграции, все чаще используются метаданные. Одним из результатов влияния разнообразия на большие данные является необходимость машиночитаемости семантики данных.

3.3.2.4 **Вариативность данных.** Вариативность данных означает изменения в скорости передачи данных, их формате/структуре, семантике и/или качестве, которые влияют на поддерживаемое приложение, аналитику или проблему. Ее влияние может заключаться в необходимости проведения реорганизации архитектур, интерфейсов, методов обработки/алгоритмов, интеграции/слияния, хранения, применимости или использования данных. Кроме того, вариативность объемов данных подразумевает необходимость увеличения или уменьшения виртуализированных ресурсов для эффективного управления дополнительной нагрузкой на обработку.

3.3.3 Ключевые параметры обработки данных

3.3.3.1 **Наука о данных.** Наука о данных изучает процесс извлечения из них знаний; используемый научный подход может заключаться либо в исследовании, либо в проверке гипотез. Наука о данных изучает полный жизненный цикл аналитики данных, в котором аналитика данных понимается согласно п.3.1.5.

3.3.3.2 **Изменчивость данных.** Изменчивость данных определяется ограниченным промежутком времени, в течение которого значения данных остаются актуальными для конкретного анализа, и выражается в виде скорости изменения во

времени. В ситуациях, когда аналитика данных проводится в режиме реального времени, немедленная обработка данных является критически необходимой для принятия решений. Наиболее очевидным образом это проявляется при работе с данными с высокой скоростью генерации, например, в случае с фондовыми рынками или в сфере телекоммуникаций. Однако данные, уже не пригодные для частной зависимой от времени аналитики ввиду устаревания, могут оставаться актуальными для других типов аналитики, не зависимых от времени.

3.3.3.3 **Достоверность данных.** Достоверность данных определяется их полнотой и точностью, для чего в отношении качества данных в профессиональном жаргоне уже длительное время существует выражение «мусор на входе – мусор на выходе». Если аналитика данных носит причинно-следственный характер, то качество каждого элемента данных является крайне важным. Если же аналитика данных проводится путем корреляции или трендирования по массивам данных большого объема, то отдельные некорректные элементы могут затеряться в общих подсчетах, и тренд все еще может быть точным.

3.3.3.4 **Выгода.** Выгода определяется степенью достижения системой обработки больших данных целей, для которых эта система создавалась.

3.3.3.5 **Визуализация данных.** Под визуализацией данных подразумевается такое их представление, которое позволяет пользователю понять информацию о демонстрируемых данных. Большие данные потребовали новых методов обработки массивов данных больших объемов, включая сбор и обобщение данных для их наибольшей наглядности. Большие данные также требуют более пристального внимания к визуальному представлению данных для лиц, ответственных за принятие решений – это необходимо для изложения результатов в доступном для понимания виде, а также для информирования об их сложности, точности и вероятностном интервале ошибок.

3.3.3.6 **Структурированные и неструктурированные данные.** Неструктурированные данные увеличиваются как в объеме, так и в значимости. Хотя реляционные базы данных обычно поддерживают эти типы элементов данных, их способность непосредственно анализировать, индексировать и обрабатывать такие типы данных, как правило, ограничена и доступна через нестандартные расширения SQL. Потребность в анализе неструктурированных данных существует уже много лет. Однако переход на парадигму больших данных привел к повышению

значимости неструктурированных данных. Также в отношении неструктурированных данных особое внимание уделяется различным новым методам разработки, которые позволят проводить анализ таких данных более эффективно.

3.3.3.7 **Масштабирование**. Большие данные подразумевают возможность расширения репозиториев данных и их обработку на параллельно работающих ресурсах - аналогичным образом сообщество моделирования, требующего ресурсоемких вычислений, массово перешло на параллельную обработку. Благодаря разработке методов взаимодействия между ресурсами, такое же масштабирование теперь доступно для приложений, использующих большое количество данных. Вертикальное масштабирование подразумевает увеличение системных параметров скорости обработки, хранения и памяти для повышения производительности. Этот подход ограничен физическими возможностями, развитие которых было описано в законе Мура, и требует все более сложных элементов (например, аппаратного и программного обеспечения), увеличивающих время и затраты на реализацию. Альтернативный метод состоит в применении горизонтального масштабирования, чтобы использовать отдельные распределенные ресурсы, объединяемые для работы в качестве единой системы. Именно горизонтальное масштабирование лежит в основе революции больших данных. Хотя методы достижения эффективной масштабируемости между ресурсами будут постоянно развиваться, эта смена парадигмы (по аналогии с предыдущим переходом на параллельную обработку в сообществе моделирования) представляет собой единовременное явление.

3.3.3.8 **Распределенная файловая система**. В распределенных файловых системах мультиструктурированные (объектные) массивы данных распределяются по вычислительным узлам кластера (кластеров) сервера. Данные могут распределяться на уровне файлов/массивов данных или – чаще всего – на уровне блоков, что позволяет нескольким узлам в кластере одновременно взаимодействовать с различными частями большого файла/массива данных. Системы больших данных часто проектируются таким образом, чтобы при распределении обработки использовать преимущества привязки данных к каждому вычислительному узлу, исключая путем этого необходимость перемещения данных между узлами. Кроме того, многие распределенные файловые системы также реализуют репликацию на уровне файлов/блоков, при которой на разных узлах компьютеров хранится несколько копий каждого файла/блока как для обеспечения

надежности/восстановления (данные не теряются при сбое узла в кластере), так и для улучшения привязки данных к вычислительным узлам. Любой тип данных и файлы любого размера могут обрабатываться без формального извлечения, преобразования и загрузки, при этом некоторые технологии работают заметно лучше с файлами большого размера.

3.3.3.9 **Распределенная обработка данных**. Популярная структура для распределенных вычислений состоит из комбинации уровня хранения и уровня обработки, которая реализует мультиклассовую модель алгоритмического программирования. Недорогие серверы потребительского уровня, поддерживающие распределенную файловую систему хранения данных, могут значительно снизить затраты на хранение вычислений для большого объема данных (например, индексация в сети). При распределенной обработке данных запрос распределен по процессорам, а результаты собираются в центральный процессор. Затем результаты обработки обычно загружаются в аналитическую среду. Для достижения эффективности, надежности, высокой доступности и отказоустойчивости системы несколько узлов (например, клиентские узлы, узлы данных, узлы-реплики) размещаются в виде архитектуры «ведущий-ведомый».

3.3.3.10 **Нереляционные базы данных**. В горизонтально масштабируемых системах данные распределяются по узлам кластера, имея при этом единую логическую структуру. Новые парадигмы базы данных нереляционной модели обычно называют системами NoSQL («не только SQL» или «не SQL»). Проблема с определением парадигмы хранения больших данных как NoSQL заключается, во-первых, в том, что она описывает хранение данных в контексте выбранного основанного на теории языка для запросов и извлечения данных, и, во-вторых, в расширении возможностей применения языков запросов, похожих на SQL, к новым нереляционным хранилищам данных. В то время как NoSQL используется настолько широко, что он будет продолжать применяться в новых моделях данных вне рамок реляционной модели, сам термин относится к базам данных, не следующим реляционной модели. Примерами моделей нереляционных баз данных являются столбец, разреженная таблица, ключ-значение, документ-ключ и графические модели.

Приложение А

(справочное)

Сквозные понятия в сфере больших данных

А.1 Общие сведения

Разработка систем больших данных имеет значение для ряда технологических сфер обсуждения и стандартизации. В данном приложении обсуждаются связи области больших данных с другими областями разработки стандартов.

А.2 Метаданные

Метаданные представляют собой описательные данные, включая, например, описание истории обработки данных. Поскольку системы больших данных спроектированы для выполнения распределенной обработки данных, в том числе тех, которые являются внешними и не находятся под контролем системы больших данных, использование метаданных становится все более важной концепцией. Поскольку большие данные повторно используются для целей, далеких от их сбора, важно, чтобы метаданные были связаны с любыми данными, доступными для других. Метаданные также включают в себя источник данных и их использование. Их можно разделить на бизнес- и технические метаданные.

А.3 Алгоритмы

При разработке алгоритмов анализа больших данных необходимо учитывать требования распределенной обработки данных, данные обычно хранились локально. В контексте больших данных алгоритмы обработки данных по узлам должны быть адаптированы к горизонтальному масштабированию, чтобы напрямую обеспечить конкретное распределение данных по узлам.

А.4 Кластерные вычисления

Кластерные вычисления относятся к распределению процессов по сети компьютеров. Компьютеры используют программное обеспечение для работы физической системы как единого целого. Если поместить уровень служб поверх физической системы, то будут достигнуты преимущества облачных вычислений.

ПРИМЕЧАНИЕ – В данном перефразированном определении кластерных вычислений под кластером понимается «комбинация набора взаимосвязанных компьютеров/серверов».

А.5 Облачные вычисления

Облачные вычисления – это одна из парадигм доступности и управления ресурсами для систем больших данных. Существует несколько ключевых характеристик, часто присущих внедрению облачных вычислений, в том числе: широкий доступ к сети, измеримое обслуживание, многопользовательский режим, самообслуживание по требованию, быстрая адаптация и масштабируемость, а также объединение ресурсов. Системы больших данных могут использовать внедрение облачных вычислений для инфраструктуры, платформ или приложений.

А.6 Безопасность данных

Системы больших данных имеют дополнительные проблемы с безопасностью из-за распределенного характера обработки данных. Дополнительные уязвимости возникают, например, при распределенном использовании и управлении физическим компьютером и сетевой инфраструктурой, а также в рамках контроля на всех уровнях программного обеспечения и сред хранения. Обычно в среде распределенной обработки данных осуществляются шифрование, маскирование и доступ на основе ролей, чтобы обеспечить комплексную защиту данных на всех уровнях, в том числе при передаче данных по сети. Некоторые примеры массивов данных, для которых требуется высокий уровень безопасности, включают в себя: конфиденциальную информацию о клиентах, информацию о продуктах, данные счетов, коммерческие данные компаний, финансовые транзакции, медицинские карты пациентов и оборонные данные.

А.7 Требования по защите конфиденциальности

Существуют законодательные и нормативные требования, которые влияют на использование личной информации и регулируют его. Все больше личной информации можно получить из сети интернет, социальных сетей, устройств слежения и т. д. В широком смысле защита конфиденциальности - это совокупность правовых и нормативных требований, которые обеспечивают право отдельных лиц на контроль не только над использованием их личной информации, но также ее достоверностью, аспектами жизненного цикла (включая принудительное удаление) и

т. д. Кроме того, ключевым правом защиты конфиденциальности является право «информированного согласия» человека в отношении использования его личной информации. Интеграция массивов данных из разнородных источников вполне может приводить к созданию наборов личной информации или получению нового способа ее использования, отличного от цели, для которой было получено осознанное согласие конкретного лица на использование такой личной информации. Поэтому любая организация, разрабатывающая и использующая системы больших данных, несет юридическую и фидуциарную ответственность за обеспечение полной поддержки и внедрения всех применимых норм по защите конфиденциальности в тех случаях, когда их деятельность связана с обработкой личной информации

A.8 SQL

SQL – это стандартный (см. серию международных стандартов ISO/IEC (ИСО/МЭК) 9075) интерактивный язык программирования, предназначенный для создания запросов, обновления и управления данными и их массивами в базе данных. SQL предназначен для манипулирования структурированными данными и предоставляет полноценную и всеобъемлющую структуру для доступа к данным, а также поддерживает широкий спектр эффективных аналитических функций. Расширения баз данных SQL поддерживают обнаружение столбцов в широком диапазоне массивов данных: не только реляционных таблиц/представлений, но также XML, JSON, пространственных объектов, объектов схожих с изображениями (больших двоичных объектов и больших символьный объектов) и семантических объектов. Системы управления данными NoSQL, предназначенные для поддержки нетабличных структурированных данных, а также неструктурированных и полуструктурированных данных, еще не сделали выбор в пользу одного общего языка доступа. Во многих вариантах реализации NoSQL приняты SQL-подобные языки, включающие некоторое подмножество стандартного SQL с расширениями, поддерживающими специфические особенности реализаций NoSQL.

A.9 Параллельные вычисления

Большие данные обычно относятся к распределенной информационно-емкой обработке данных узлами кластера. Сообщество моделирования уже много лет разрабатывает методы информационно-емкой обработки большими кластерами вычислительных узлов. Учитывая, что оба подхода представляют собой крайние

случаи для высокомасштабированных вычислений и анализа данных, технологии обоих подходов будут использоваться для спектра возможностей, требующих как ресурсоемких, так и информационно-емких вычислений.

А.10 Интернет вещей

Одновременно с созданием все большего и большего количества данных, создаются вычислительные системы, способные эти данные анализировать. Пользователи хотят использовать объем данных, доступных с различных сенсоров и других источников данных. Это обеспечивает эффективную предсказательную аналитику данных для управления и контроля сетевых решений. Типичные технологические достижения в области сенсоров, а также развертывание IPV6 для обеспечения подключения этих устройств к сети интернет, создают потребность в системе больших данных, которая сможет обрабатывать потоковые данные, обладающие высокой скоростью генерации, из нескольких источников. Это отличается от систем с крупными объемами больших данных, которые обычно запускают пакетные задания на относительно небольшом количестве больших массивов данных. Данная разница в характеристиках массивов данных оказывает прямое влияние на архитектуру и методы, используемые для анализа данных.

А.11 Языки программирования

Анализ расширенных данных с использованием статистических вычислений является фундаментальным подходом к концепции больших данных. Пользователи могут разрабатывать системы аналитики больших данных с использованием языков программирования общего назначения. Потребности в распределенной обработке данных привели к появлению ряда новых языков программирования и запросов, подходящих для разработки систем больших данных, а также новых процессов. Языки программирования (см. примечание 1), как правило, имеют доступные платформы, библиотеки и средства динамической поддержки для обеспечения эффективной обработки больших данных с использованием параллельных вычислений и хранения. Среди новых процессов – распределение-сборка данных для их распределенной обработки.

ПРИМЕЧАНИЕ – Примеры языков включают в себя R, Python, Scala, Java и т. д.

Алфавитный указатель терминов на русском языке

Приложение ДА

(справочное)

Сведения о соответствии ссылочных международных стандартов национальным стандартам

Таблица ДА.1

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO/IEC 2382:2015	MOD	ГОСТ 33707-2016 (ISO/IEC 2382:2015) Информационные технологии (ИТ). Словарь
ISO 9075 (all parts)	–	*
ISO/IEC 11404	–	*
ISO/IEC 17788:2014	IDT	ГОСТ ISO/IEC 17788-2016 Информационные технологии (ИТ). Облачные вычисления. Общие положения и терминология
ISO/IEC 19784-4:2011	IDT	ГОСТ Р ИСО/МЭК 19784-4-2014 Информационные технологии (ИТ). Биометрия. Биометрический программный интерфейс. Часть 4. Интерфейс поставщика функции биометрического датчика
* Соответствующий международный стандарт отсутствует. Примечание – В настоящей таблице использованы следующие условные обозначения степени соответствия стандартов: - IDT - идентичные стандарты; - MOD - модифицированные стандарты.		

Библиография

[1] ГОСТ 33707-2016 (ISO/IEC 2382:2015) Информационные технологии (ИТ). Словарь

[2] ГОСТ ISO/IEC 17788-2016 Информационные технологии (ИТ). Облачные вычисления. Общие положения и терминология

[3] ГОСТ Р ИСО/МЭК 19784-4-2014 Информационные технологии (ИТ). Биометрия. Биометрический программный интерфейс. Часть 4. Интерфейс поставщика функции биометрического датчика

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии (ИТ), данные, большие данные, аналитика данных, база данных, модель данных, наука о данных, массив данных, тип данных, вариативность данных, разнообразие данных, скорость обработки данных, достоверность данных, изменчивость данных, объем данных, распределенная обработка данных, неструктурированные данные, частично структурированные данные, потоковые данные

Руководитель организации-разработчика

Директор
Национального центра цифровой экономики
МГУ имени М.В. Ломоносова Т.В. Ершова

Руководитель разработки

Председатель совета директоров
АНО «Институт развития
информационного общества» Ю.Е. Хохлов