ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСТ Р ИСО/МЭК 20546— 2021

Информационные технологии

БОЛЬШИЕ ДАННЫЕ

Обзор и словарь

(ISO/IEC 20546:2019, IDT)

Издание официальное

Москва
Стандартинформ
2021

Предисловие

1 ПОДГОТОВЛЕН Федеральным государственным бюджетным образовательным учреждением высшего образования «Московский государственный университет имени М.В. Ломоносова» (МГУ имени М.В. Ломоносова) в лице Научно-образовательного центра компетенций в области цифровой экономики МГУ и Автономной некоммерческой организацией «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии стандарта, указанного в пункте 4

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 13 июля 2021 г. № 632-ст

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

Дополнительные сноски в тексте стандарта, выделенные курсивом, приведены для пояснения текста стандарта

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 порядке, отражающем систему понятий данной области знания.

Для каждого понятия установлен один стандартизованный термин.

В стандарте приводятся эквиваленты стандартизованных терминов на английском (en) 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. Overview and vocabulary

Дата введения — 2021—11—01

1 Область применения

Настоящий стандарт содержит набор терминов и определений, необходимых для улучшения информационного взаимодействия и формирования русскоязычных понятий в области информационных технологий и больших данных. Он обеспечивает терминологическую основу для стандартов, связанных с большими данными.

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

2 Нормативные ссылки

Нормативные ссылки в настоящем стандарте отсутствуют.

3 Термины и определения

В настоящем стандарте применены следующие термины с соответствующими определениями.

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 практически полезных 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), отличающиеся главным образом такими характеристиками, как объем, разнообразие, скорость обработки и/или вариативность, которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа.

Примечание — Термин «большие данные» широко применяется в различных значениях, например в качестве наименования технологии масштабирования, используемой для обработки больших массивов данных.

Издание официальное

3.1.3 **облачные вычисления** (cloud computing): Парадигма для предоставления возможности сетевого доступа к масштабируемому и эластичному пулу общих физических или виртуальных ресурсов с предоставлением самообслуживания и администрированием по требованию.

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

[ИСО/МЭК 17788:2014, 3.2.5]

3.1.4 **кластер** (в распределенной обработке данных) (cluster): Совокупность функциональных устройств, находящихся под общим управлением.

[ИСО/МЭК 2382:2015, 4.496]

3.1.5 **данные** (data): Представление информации в формальном виде, пригодном для передачи, интерпретации или обработки.

Примечание — Данные могут быть обработаны автоматически или вручную.

[ИСО/МЭК 2382:2015, 4.259]

3.1.6 **аналитика данных** (data analytics): Составное понятие, охватывающее получение, сбор, проверку и обработку данных (3.1.9), включая их количественную оценку, визуализацию и интерпретацию.

Примечание — Аналитика данных используется для представления объектов, описываемых данными (3.1.5), с целью прогнозирования конкретных ситуаций и формирования пошаговых рекомендаций при решении задач. Закономерности, полученные посредством аналитики, используются в различных целях, таких как принятие решений, проведение исследований, обеспечение устойчивого развития, проектирование, планирование и т. д.

3.1.7 **база данных** (database): Совокупность данных (3.1.5), организованная в соответствии с концептуальной структурой, в которой описываются характеристики этих данных и взаимосвязи между представляемыми ими сущностями для одной или нескольких областей применения.

[ИСО/МЭК 2382:2015, 2121413]

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

[ИСО/МЭК 2382:2015, 2125519]

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

Примечания

1 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 Термин «обработка данных» не должен использоваться в качестве синонима для термина «обработка информации».

[ИСО/МЭК 2382:2015, 2121276]

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.

[Адаптировано из ИСО 19115-2:2009, 4.7]

3.1.12 **тип данных** (data type, datatype): Совокупность объектов данных (3.1.5) установленной структуры и набора допустимых операций над этими объектами.

Примечания

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

2 Если толкование не вызывает сомнений, то вместо термина «тип данных» может быть использован термин «тип».

3 Тип данных: определение и термины, стандартизованные ИСО/МЭК [ИСО/МЭК 2382-15:1999].

[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 — Данное понятие отражает нерегулярность и разнородность структур данных, навигации по структурам, запросов и типов данных.

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

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

Примечание — Под достоверностью данных понимают описательные 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 их изменения с течением времени.

[ИСО/МЭК 2382:2015, 2121506]

3.1.18 **объем данных** (data volume): Количественная характеристика данных (3.1.5), влияющая на выбор ресурсов для вычислений и хранения, а также на управление данными в процессе обработки.

Примечание — Объем данных становится важным при работе с большими массивами данных (3.1.11).

3.1.19 **распределенная обработка данных** (distributed data processing): Обработка данных, при которой выполнение операций распределено по узлам вычислительной сети.

[ИСО/МЭК 2382:2015, 4.1166]

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

3.1.21 **файл** (file): Поименованная совокупность записей, рассматриваемая как единое целое.

[ИСО/МЭК 2382:2015, 4.1470]

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

Примечание — См. распределение—сборка (3.2.33)[1)].

3.1.23 **горизонтальное масштабирование** (horizontal scaling): Формирование единого логического целого путем соединения нескольких аппаратно-программных комплексов.

Примечания

1 Примером горизонтального масштабирования является повышение производительности распределенной обработки данных путем добавления узлов в кластере с целью подключения (привлечения) 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 описание, а также сведения о владельце данных, путях доступа к ним, правах доступа и изменчивости данных (3.1.17).

[ИСО/МЭК 2382:2015, 2121505]

3.1.25 **нереляционная база данных** (non-relational database): База данных (3.1.7), не соответствующая 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 соответствующая реляционной модели (3.1.31) хранения и обработки данных (3.1.5).

3.1.27 **параллельность** (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 — Параллельная передача битов компьютерного слова по линиям внутренней шины.

[ИСО/МЭК 2382:2015, 2121654]

---

1) *Согласно оригиналу.*

3.1.28 **частично структурированные данные** (partially structured data): Данные (3.1.5), в которых присутствует определенная степень организации.

Примечания

1 Частично структурированные данные в практической деятельности часто называют полуструктурированными.

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

[ИСО/МЭК 2382:2015, 2121654]

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

[ИСО/МЭК 2382:2015, 2121473]

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

[ИСО/МЭК 2382:2015, 2121470]

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

[ИСО/МЭК 2382:2015, 2121469]

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

Примечание — См. распределение—сборка (3.2.33)[1)].

3.1.33 **распределение—сборка** (scatter-gather): Вид обработки больших массивов данных (3.1.11), при которой 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.

[ИСО/МЭК 19784-4:2011, 4.4]

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

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 ИСО/МЭК 9075.

Примечание — Аббревиатура SQL иногда расшифровывается как «язык структурированных запросов» (Structured Query Language), но это название не используется в серии стандартов 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 Примером неструктурированных данных является произвольный текст.

3.1.38 **вертикальное масштабирование** (vertical scaling): Повышение производительности обработки данных за счет улучшения характеристик процессоров, памяти, хранения или коннективности.

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

[1)] *Согласно оригиналу.*

3.2 Сокращения

JSON — обозначение объектов Javascript;
PII — персональные данные;
XML — расширяемый язык разметки.

4 Ключевые характеристики больших данных

4.1 Общие сведения

При выборе системы больших данных необходимо руководствоваться четырьмя характеристиками — объемом, скоростью обработки, разнообразием и вариативностью данных (см. 4.2.4). Управление этими характеристиками определяется средствами обработки в соответствии с описанием в 4.2.

4.2 Ключевые характеристики данных

4.2.1 **Объем данных.** Объем данных представляет собой определенное количество данных, доступных для анализа с целью извлечения полезной информации. Одним из основных факторов развития технологий обработки больших данных стал рост объемов данных, как следствие интернет-активности.

4.2.2 **Скорость обработки данных.** Скорость обработки данных — это скорость потока создания, хранения, анализа и визуализации данных. Скорость обработки больших данных означает необходимость обработки большого количества данных за короткий промежуток времени. В качестве примеров работы с данными с высокой скоростью обработки обычно приводят средства обработки потоковых данных.

4.2.3 **Разнообразие данных.** Свойство разнообразия данных отражает необходимость анализа данных разных типов из различных предметных областей. Как правило, проблема разнообразия данных решалась посредством их преобразования или проведения предварительного анализа с целью выявления свойств, позволяющих объединить их с другими данными. Более широкий диапазон форматов данных, логических моделей, временных шкал и семантики, которые предпочтительно использовать при аналитике, усложняет интеграцию разнообразных данных. В качестве средства, способствующего интеграции, все чаще используются метаданные. Одним из результатов влияния разнообразия на большие данные является необходимость представления семантики данных в машиночитаемом виде.

4.2.4 **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. Влияние вариативности может заключаться в необходимости проведения реорганизации архитектур, интерфейсов, методов обработки/алгоритмов, интеграции/слияния, хранения, применимости или использования данных. В дополнение следует отметить, что вариативность объемов данных подразумевает необходимость увеличения или уменьшения виртуализированных ресурсов для эффективного управления дополнительной нагрузкой на обработку.

4.3 Ключевые характеристики обработки данных

4.3.1 **Наука о данных.** Наука о данных изучает процесс извлечения из них знаний. Используемый научный подход может заключаться либо в проведении исследований, либо в проверке гипотез. Наука о данных изучает полный жизненный цикл аналитики данных, в котором аналитика данных понимается согласно 3.1.6.

4.3.2 **Изменчивость данных.** Изменчивость данных связана с ограниченным промежутком времени, в течение которого значения данных остаются актуальными для конкретного анализа, и определяется динамикой изменений.

В тех 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 с данными, связанными с фондовыми рынками или телекоммуникациями. Однако данные, не пригодные для специфического, чувствительного к временным рамкам анализа ввиду устаревания, могут оставаться актуальными для других типов аналитики, не зависимых от времени.

4.3.3 **Достоверность данных.** Достоверность данных определяется их полнотой и точностью, в связи с чем для обозначения качества данных в профессиональном жаргоне длительное время су-

ществует выражение «мусор на входе — мусор на выходе». Если аналитика данных направлена на установление причинно-следственных связей, то качество каждого элемента является крайне важным. Если аналитика осуществляется путем выявления корреляций или трендов в больших массивах данных, то отдельные некорректные элементы могут быть утеряны при общих подсчетах, но тренд может оставаться точным.

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

4.3.5 **Визуализация данных.** Под визуализацией данных подразумевается такое их представление, которое позволяет пользователю извлечь из них информацию. 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 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.

4.3.6 **Структурированные и неструктурированные данные.** Постоянно увеличиваются как объемы, так и значение неструктурированных данных. 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.

4.3.7 **Масштабирование.** Большие данные подразумевают возможность расширения репозиториев данных и их обработку на параллельно работающих ресурсах — аналогичным образом сообщество специалистов, использующих моделирование, требующее 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 описано в законе Мура, и требует все более сложных элементов (например, аппаратного и программного обеспечения), приводящих к росту материальных и временных затрат 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 специалистов, используемых моделирование) представляет собой единовременное явление.

4.3.8 **Распределенная файловая система.** В распределенных файловых системах мультиструктурированные (объектные) массивы данных распределяются по вычислительным узлам кластера(ов) серверов. 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 необходимость перемещения данных между узлами. Кроме того, во многих распределенных файловых системах также реализована репликация 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 эффективнее с файлами большого размера.

4.3.9 **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 распределен-

ной 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».

4.3.10 **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 NoSQL («не только SQL» или «не SQL»). Проблема с определением парадигмы хранения больших данных как NoSQL заключается, во-первых, в описании хранения данных на теоретико-множественном языке 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 будет 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

Развитие систем больших данных оказывает влияние на дискуссии и процессы стандартизации в других технологических областях. 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. Системы больших данных спроектированы для выполнения распределенной обработки данных, в том числе тех, которые являются внешними и не находятся под контролем системы больших данных, поэтому использование метаданных становится все более значимой концепцией. Большие данные повторно используются для целей, не связанных с целями, для которых они собирались, поэтому важно, чтобы любые данные, доступ к которым предоставляется другим сторонам, были снабжены адекватными метаданными. Метаданные также включают в себя сведения об источниках данных и об использовании данных. Их можно разделить на бизнес- и технические метаданные.

А.3 Алгоритмы

При разработке алгоритмов анализа больших данных необходимо учитывать требования распределенной обработки, поскольку ранее данные обычно хранились на локальных устройствах. Алгоритмы обработки больших данных в узлах должны быть адаптированы к горизонтальному масштабированию, чтобы явно учитывать распределение данных по узлам.

А.4 Кластерные вычисления

Кластерные вычисления относятся к распределению процессов по компьютерной сети. Для работы аппаратной среды как единого целого используется специализированное программное обеспечение. Если поместить уровень служб поверх аппаратной среды, 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 — одна из парадигм доступности и управления ресурсами для систем больших данных. Существует несколько ключевых характеристик, часто присущих применению облачных вычислений, в том числе: широкополосный доступ, измеримое обслуживание, многопользовательский режим, самообслуживание по требованию, быстрая адаптация и масштабируемость, а также объединение ресурсов. Облачные вычисления для инфраструктуры, платформ или приложений могут применяться при формировании системы больших данных.

А.6 Безопасность данных

Системы больших данных из-за распределенного характера обработки имеют дополнительные проблемы с безопасностью. Дополнительные уязвимости возникают, например, при распределенном использовании и управлении физической компьютерной и сетевой инфраструктурами, а также при контроле доступа на каждом слое программного обеспечения и системы хранения. Обычно в среде распределенной обработки данных осуществляются шифрование, маскирование и управление доступом на основе ролей для обеспечения всесторонней защиты данных на всех слоях, включая передачу данных по сети. Некоторые примеры массивов данных, для которых обязателен высокий уровень безопасности, включают конфиденциальную информацию о клиентах, сведения о продуктах, коммерческие сведения компаний, данные счетов и финансовые транзакции, медицинские записи пациентов, а также сведения, относящиеся к национальной обороне и безопасности.

А.7 Требования по защите персональных данных

Существуют законодательные и нормативные требования, которые влияют на использование персональных данных и регулируют его. Все больше персональных данных можно получить из сети Интернет, социальных сетей, устройств слежения и т. д. В широком смысле защита персональных данных — это совокупность правовых и нормативных требований, которые обеспечивают право отдельных лиц на контроль не только над использованием их персональных данных, но также их достоверностью, аспектами жизненного цикла (включая принудительное удаление) и т. д. Кроме того, ключевым правом защиты персональных данных является право «информированного согласия» человека в отношении использования его персональных данных. Интеграция массивов данных из разнородных источников может приводить к созданию наборов персональных данных или получению нового способа их

использования, отличного от цели, для которой получено осознанное согласие конкретного лица на использование таких персональных данных. 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 персональных данных в тех случаях, когда их деятельность связана с обработкой персональных данных.

А.8 SQL

SQL — это стандартный (см. серию стандартов ИСО/МЭК 9075) интерактивный язык программирования, предназначенный для создания запросов, обновления и управления данными и их массивами в базе данных. SQL предназначен для работы со 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 символьных 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.

А.9 Параллельные вычисления

Большие данные обычно связаны с распределенной интенсивной обработкой данных в узлах кластера. Сообщество специалистов в области моделирования уже много лет разрабатывает методы интенсивного использования компьютерных вычислений в больших вычислительных кластерах. Учитывая, что оба подхода представляют собой крайние случаи для крупномасштабных вычислений и анализа данных, технологии обоих подходов будут использоваться для спектра возможностей, требующих как интенсивных компьютерных вычислений, так и интенсивной обработки данных.

А.10 Интернет вещей

Одновременно с увеличением объема данных создаются вычислительные системы, способные эти данные анализировать. Пользователи предпочитают использовать объем данных, доступных с различных сенсоров и других источников, что обеспечивает эффективную предсказательную аналитику для управления и контроля сетевых решений. Технологические достижения в области сенсоров, а также развертывание протокола IPV6 для обеспечения интернет-коннективности этих устройств порождают потребность в системах больших данных, которые могут обрабатывать потоковые данные из нескольких источников, обладающих высокой скоростью генерации. Подобные системы отличаются от систем, создаваемых для пакетной обработки малого числа больших массивов данных. Различие в характеристиках массивов данных оказывает прямое влияние на архитектуру систем и методы анализа данных.

А.11 Языки программирования

Анализ расширенных данных с использованием статистических вычислений является фундаментальным методом в парадигме больших данных. Системы аналитики больших данных могут разрабатываться с использованием базовых языков программирования. Потребности в распределенной обработке данных привели к появлению новых языков программирования, языков запросов и процессов обработки, пригодных для создания систем больших данных. Языки программирования (см. примечание), как правило, имеют общедоступные среды разработки, библиотеки и среды выполнения 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.

Библиография

[1] ISO/IEC 2382:2015, *Information technology — Vocabulary*

[2] ISO 9075 (all parts), *Information technology — Database languages — SQL*

[3] ISO/IEC 11404, *Information technology — General-Purpose Datatypes (GPD)*

[4] ISO/IEC 17788:2014, *Information technology — Cloud computing — Overview and vocabulary*

[5] ISO/IEC 19784-4:2011, *Information technology — Biometric application programming interface — Part 4: Biometric sensor function provider interface*

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

Технический редактор *И.Е. Черепкова*
Корректор *Е.Д. Дульнева*
Компьютерная верстка *Е.А. Кондрашовой*

Сдано в набор 13.07.2021. Подписано в печать 15.07.2021. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 1,86. Уч.-изд. л. 1,68.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении во ФГУП «СТАНДАРТИНФОРМ»
для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru