**Ключевые термины по теме «Обработка больших данных и извлечение знаний»**

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

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

1. **массив** **данных** (data set, dataset): идентифицируемая совокупность данных (1), к которой можно получить доступ или скачать в одном или нескольких форматах.

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

1. **большие** **данные** (big data): большие массивы данных (2), – главным образом, по таким характеристикам данных (1), как объем, разнообразие, скорость обработки и/или вариативность, – которые требуют использования технологии масштабирования для эффективного хранения, обработки, управления и анализа.
2. **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]

1. **модель** **данных** (data model): схема структурирования данных (1) в базе данных (4) 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]

1. **обработка** **данных** (data processing): систематическое выполнение операций с данными (1).
2. **аналитика** **данных** (data analytics): составное понятие, состоящее из получения, сбора, проверки и обработки данных (6), включая их количественную оценку, визуализацию и интерпретацию.
3. **наука** **о** **данных** (data science): извлечение практических знаний из данных (1) посредством исследования или создания и проверки гипотез.
4. **тип** **данных** (data type, datatype): определенный массив объектов данных (1) конкретной структуры данных и набор допустимых операций, в рамках которых эти объекты данных выступают в роли операндов при выполнении любой из этих операций.
5. **вариативность данных** (data variability): изменения в скорости передачи, формате или структуре, семантике или качестве массива данных (2).
6. **разнообразие данных** (data variety): диапазон форматов, логических моделей, временных шкал и семантики массива данных (2).
7. **скорость обработки данных** (data velocity): скорость потока, с которой данные (1) создаются, передаются, хранятся, анализируются или визуализируются.
8. **достоверность данных** (data veracity): полнота и/или точность данных (1).
9. **изменчивость данных** (data volatility): характеристика данных (1), относящаяся к скорости изменения этих данных с течением времени.

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

1. **объем данных** (data volume): степень количества данных (1), оказывающая влияние на ресурсы для вычислений и хранения, а также на управление ими в процессе обработки данных.
2. **распределенная обработка данных** (distributed data processing): обработка данных (6), в которой выполнение операций распределено между узлами компьютерной сети. [ИСТОЧНИК: Международный стандарт ISO/IEC (ИСО/МЭК) 2382:2015, 18.01.08]
3. **кластер** (cluster): (распределенная обработка данных) набор функциональных блоков под общим контролем.

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

1. **сборка** (gather): объединение результатов из нескольких узлов в кластере
2. **горизонтальное масштабирование** (horizontal scaling): формирование единого логического блока путем соединения нескольких аппаратных и программных средств.
3. **метаданные** (metadata): данные (1) 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]

1. **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.
2. **распределение** (scatter): Распределение обработки по нескольким узлам в кластере (18).
3. **распределение-сборка** (scatter-gather): вид обработки больших массивов данных (2), где необходимые вычисления разделяются и распределяются по нескольким узлам в кластере, а общий результат формируется путем объединения результатов от каждого узла.
4. **потоковые данные** (streaming data): данные (1), передаваемые через интерфейс от непрерывно работающего источника.

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

1. **структурированные данные** (structured data): данные (1), организованные на основе предопределенного (применимого) набора.
2. **неструктурированные данные** (unstructured data): данные (1), характеризуемые отсутствием какой-либо структуры, кроме структуры на уровне записи или файла.
3. **вертикальное масштабирование** (vertical scaling): повышение производительности обработки данных за счет улучшения процессоров, памяти, хранилища или связи.
4. **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.