Большие данные (Big data)

> Loginom: Руководство пользователя

В узком смысле — массивы структурированных, слабоструктурированных и неструктурированных данных, объемы которых настолько велики, что их обработка традиционными средствами становится неэффективной или вообще невозможной.

В широком смысле — комплекс средств и методов для обработки и анализа массивов данных, подпадающих под определение больших данных.

Изначально с большими данными связывали три ключевых концепции (правило «трех V»):

- **Объем (volume)**. Данные в компании накапливаются из множества источников в громадном объеме.
- **Скорость роста (velocity)**. Быстрое возрастание объемов данных. Особенно характерно для компаний в области сетевой торговли и электронной коммерции, где ежедневно могут генерироваться сотни терабайт данных.
- **Многообразие (variety)**. Данные из входного потока могут быть разнообразных форматов (таблицы, текст, видео, аудио и пр.), а также быть структурированными и неструктурированными.

Постепенно правило «трех V» обогатилось дополнительными элементами и трансформировалось в: «четыре V» (veracity — **достоверность**), «пять V» (viability — **жизнеспособность** и value — **ценность**) и «семь V» (variability — **переменчивость** и visualization — **визуализация**).

В настоящее время понятие «большие данные» связано с использованием предсказательной и поведенческой аналитики и других направлений анализа данных с целью извлечения знаний из огромных массивов данных.

Главными проблемами, с которыми приходится сталкиваться при работе с большими данными, являются возрастание вычислительных затрат — как в плане времени, так и требуемых объемов памяти. Отсюда вытекают задачи оптимизации размещения данных в оперативной памяти, количества обращений к диску и числа проходов по данным.

Если обработка данных невозможна на одном компьютере, то ее алгоритм можно разделить на части и попытаться выполнить на нескольких машинах. Эта идея послужила толчком для появления и развития методологий и инструментов распределенной обработки, например, MapReduce, HDFS, Hive.

Для снижения количества итераций и/или проходов по набору данных при работе аналитических алгоритмов используются их различные вероятностные модификации. Примером такого алгоритма является оптимальное зависимое от данных хеширование для приближенного поиска ближайших соседей.

С большими данными сталкиваются во многих сферах: науке, электронной коммерции, телекоммуникациях, финансовом секторе. Кроме того, для решения бизнес-задач можно привлекать данные из сторонних источников.

Например, информация о пользовательской активности, связях и интересах из социальных сетей может использоваться для обогащения данных при персонализации маркетинговых предложений или при прогнозе платежеспособности заемщика в скоринге.

Термин «большие данные» получил широкое использование начиная с 1990-х годов, а его популяризацию связывают с именем John R. Mashey.