**Большие данные**

**Большие данные (**[**англ.**](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA)**big data**) — серия подходов, инструментов и методов обработки структурированных и [неструктурированных данных](https://ru.wikipedia.org/w/index.php?title=%D0%9D%D0%B5%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%82%D1%83%D1%80%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5&amp;action=edit&amp;redlink=1) огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам [вычислительной сети](https://ru.wikipedia.org/wiki/%D0%9A%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BD%D0%B0%D1%8F_%D1%81%D0%B5%D1%82%D1%8C), сформировавшихся в конце [2000-х годов](https://ru.wikipedia.org/wiki/2000-%D0%B5), альтернативных традиционным [системам управления базами данных](https://ru.wikipedia.org/wiki/%D0%A1%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0_%D1%83%D0%BF%D1%80%D0%B0%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F_%D0%B1%D0%B0%D0%B7%D0%B0%D0%BC%D0%B8_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85) и решениям класса [Business Intelligence](https://ru.wikipedia.org/wiki/Business_Intelligence).

**Общая архитектура, характерная для Big Data**:

1. Сбор данных. Источники данных (информации) сами по себе не входят в архитектуру Big Data, но в нее включаются программные и технические средства, способные осуществлять сбор. Механизмы сбора можно классифицировать следующим образом:
   * сбор структурированных данных (различные базы данных, существующие архивы данных, внутренние системы предприятий и т.д.)
   * сбор неструктурированных данных (данные GPS, аудио- и видеофайлы, текстовые файлы, аналоговые источники информации и т.д.)
   * сбор частично структурированных данных (данные журналов событий внутренних систем, сетевых служб, XML-данные и т.д.)
2. Хранилище данных. Все собранные данные распределяются на хранение и, в зависимости от типа данных, оказываются в распределенных/нераспределенных хранилищах или фиксируются в отдельных журналах записи событий
3. Преобразование данных. Перед передачей данных на стадию обработки они должны быть преобразованы в понятный для программ формат с помощью инструментов импорта/экспорта
4. Обработка данных. На данном этапе происходит объединение всех собранных данных. Обработка может проходить пакетами или в режиме реального времени
5. Анализ данных. Инструменты, используемые на данном этапе, зависят от целей пользователя. Для решения отдельных задач могут быть созданы самостоятельные алгоритмы анализа, утилиты или использованы стандартные.
6. Вывод данных. Результаты анализа должны быть представлены в формате, удобном для восприятия пользователем. Это могут быть таблицы, диаграммы, машинописный текст и т.д. – должна быть проведена визуализация результатов.

Исходя из определения **Big Data**, можно сформулировать основные **принципы работы** с такими данными:

1. **Горизонтальная масштабируемость**. Любая система, которая подразумевает обработку больших данных, должна быть расширяемой, так как количество данных может быть очень большим
2. **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.
3. **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 Big Data-решений является принцип локальности данных – по возможности обрабатываем данные на той же машине, на которой их храним.

В качестве определяющих характеристик для больших данных традиционно **выделяют «три V**»:

* объём ([англ.](https://ru.wikipedia.org/wiki/%D0%90%D0%BD%D0%B3%D0%BB%D0%B8%D0%B9%D1%81%D0%BA%D0%B8%D0%B9_%D1%8F%D0%B7%D1%8B%D0%BA) *volume*, в смысле величины физического объёма)
* скорость (*velocity* в смыслах как скорости прироста, так и необходимости высокоскоростной обработки и получения результатов)
* многообразие (*variety*, в смысле возможности одновременной обработки различных типов, структурированных и полу структурированных данных)

**Методы и техники анализа**, применимые к большим данным:

* методы класса [Data Mining](https://ru.wikipedia.org/wiki/Data_Mining): обучение ассоциативным, классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным), [кластерный анализ](https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7), [регрессионный анализ](https://ru.wikipedia.org/wiki/%D0%A0%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D0%BE%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7);
* [краудсорсинг](https://ru.wikipedia.org/wiki/%D0%9A%D1%80%D0%B0%D1%83%D0%B4%D1%81%D0%BE%D1%80%D1%81%D0%B8%D0%BD%D0%B3) — категоризация и обогащение данных силами широкого, неопределённого круга лиц, привлечённых на основании публичной оферты, без вступления в трудовые отношения;
* смешение и интеграция данных — набор техник, позволяющих интегрировать разнородные данные из разнообразных источников для возможности глубинного анализа, в качестве примеров таких техник, составляющих этот класс методов приводятся [цифровая обработка сигналов](https://ru.wikipedia.org/wiki/%D0%A6%D0%B8%D1%84%D1%80%D0%BE%D0%B2%D0%B0%D1%8F_%D0%BE%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D1%81%D0%B8%D0%B3%D0%BD%D0%B0%D0%BB%D0%BE%D0%B2) и [обработка естественного языка](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%80%D0%B0%D0%B1%D0%BE%D1%82%D0%BA%D0%B0_%D0%B5%D1%81%D1%82%D0%B5%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%BE%D0%B3%D0%BE_%D1%8F%D0%B7%D1%8B%D0%BA%D0%B0)
* [машинное обучение](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5), включая [обучение с учителем](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D1%81_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D0%B5%D0%BC) и [без учителя](https://ru.wikipedia.org/wiki/%D0%9E%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B1%D0%B5%D0%B7_%D1%83%D1%87%D0%B8%D1%82%D0%B5%D0%BB%D1%8F), а также использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей, [искусственные нейронные сети](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BD%D0%B5%D0%B9%D1%80%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D1%81%D0%B5%D1%82%D1%8C), [сетевой анализ](https://ru.wikipedia.org/wiki/%D0%A1%D0%B5%D1%82%D0%B5%D0%B2%D0%BE%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7), [оптимизация](https://ru.wikipedia.org/wiki/%D0%9E%D0%BF%D1%82%D0%B8%D0%BC%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F_(%D0%BC%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0))
* [распознавание образов](https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BE%D1%80%D0%B8%D1%8F_%D1%80%D0%B0%D1%81%D0%BF%D0%BE%D0%B7%D0%BD%D0%B0%D0%B2%D0%B0%D0%BD%D0%B8%D1%8F_%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BE%D0%B2)
* [прогнозная аналитика](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%B3%D0%BD%D0%BE%D0%B7%D0%BD%D0%B0%D1%8F_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D1%82%D0%B8%D0%BA%D0%B0)
* [имитационное моделирование](https://ru.wikipedia.org/wiki/%D0%98%D0%BC%D0%B8%D1%82%D0%B0%D1%86%D0%B8%D0%BE%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BC%D0%BE%D0%B4%D0%B5%D0%BB%D0%B8%D1%80%D0%BE%D0%B2%D0%B0%D0%BD%D0%B8%D0%B5)
* [пространственный анализ](https://ru.wikipedia.org/w/index.php?title=%D0%9F%D1%80%D0%BE%D1%81%D1%82%D1%80%D0%B0%D0%BD%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D1%85&action=edit&redlink=1) — класс методов, использующих [топологическую](https://ru.wikipedia.org/wiki/%D0%A2%D0%BE%D0%BF%D0%BE%D0%BB%D0%BE%D0%B3%D0%B8%D1%8F), [геометрическую](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BE%D0%BC%D0%B5%D1%82%D1%80%D0%B8%D1%8F) и [географическую](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BE%D0%B3%D1%80%D0%B0%D1%84%D0%B8%D1%8F) информацию в данных;
* [статистический анализ](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7)
* визуализация аналитических данных — представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа.

**Плюсы** технологии “Big Data”:

* Скорость обработки и принятия решений
* Надежная база для бизнес-решений
* Эффективность результатов
* Выявление закономерностей
* Понятные метрики
* Повышенная производительность
* Снижение затрат
* Улучшение обслуживания клиентов
* Большие данные как средства для инноваций и перестройки рынка

**Минусы** технологии “Big Data”:

* Необходимость реорганизации бизнес-процессов
* Угроза конфиденциальности и контроль
* Недостаток специалистов
* Качество данных
* Потребность в оборудовании
* Сложность интеграции