Что же такое **BIG DATA?**

- Big Data — это наборы данных такого объема, что традиционные инструменты не способны осуществлять их захват, управление и обработку за приемлемое для практики время.

- Технология Big Data предоставляет услуги, помогающие раскрыть коммерческий потенциал мегамассивов данных за счет поиска ценных закономерностей и фактов путем объединения и анализа больших объемов данных.

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

Интернет и мобильные технологии

Twitter 175 млн твит сообщений в день

Facebook 300 млн фото загружаемых ежедневно

Google 24PB ежедневно

AT&T передает 30Pb в день

Walmart более 1 млн продаж в час

Объем данных, переданных/полученных на мобильные устройства, — 1,3 эксабайт

Основные технологии анализа в **BigData**

- MapReduce - это фреймворк для вычисления некоторых наборов распределенных задач с использованием большого количества компьютеров (называемых «нодами»), образующих кластер, разработанный компанией Google.
- Hadoop - набор утилит, библиотек и программный каркас для разработки и выполнения распределённых программ, работающих на кластерах из сотен и тысяч узлов.
- NoSql - ряд подходов, направленных на реализацию хранилищ баз данных, имеющих существенные отличия от моделей, используемых в традиционных реляционных СУБД с доступом к данным средствами языка SQL. Применяется к базам данных, в которых делается попытка решить проблемы масштабируемости и доступности за счёт атомарности и согласованности данных

Методы анализа используемые в **BigData**

Уникальность подхода больших данных заключается в агрегировании огромного объема неструктурированной информации из разных источников в одном месте.

- Классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным)
- Кластерный анализ
- Регрессионный анализ
- Рекомендательные системы
- Искусственные нейронные сети, в том числе генетические алгоритмы

Способы повышения производительности

Производительность при обработке больших объемов данных можно повысить различными способами:

Оборудование: многопроцессорные системы, ОЗУ большой емкости, RAID-массивы...

Базы данных: «тяжелые» СУБД, разбиение на разделы, оптимальное индексирование...

Аналитическая платформа: параллельная обработка, кэширование данных, комбинирование простых и сложных моделей...

Исходная информация: репрезентативные выборки, сегментирование данных, группировка...

Алгоритмы: масштабируемые алгоритмы, комитеты моделей, иерархические модели...

Самые продвинутые отрасли BigData

01 Маркетинг

- Сегментация рынка
- Моделирование приобретения и оттока клиентов
- Рекомендательные системы
- Анализ соц.медиа

02 Финансы

- Детектирование аномального поведения
- Анализ кредитных рисков
- Страховое моделирование

03 Медицина

- Генетический анализ
- Анализ клинических испытаний
- Экспертные системы