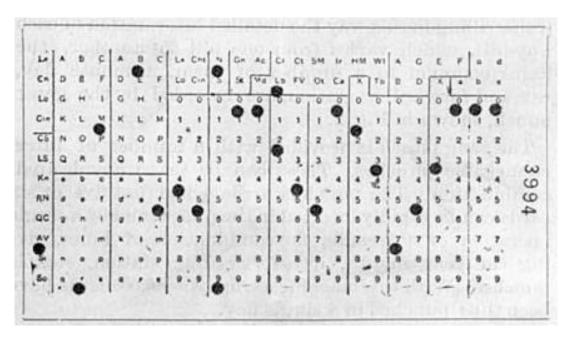
BigData, типы данных и инструменты

Лекция 1

Начало работы с BigData

• В 1890 году при переписи впервые использовали электрическую табулирующую машину Германа Холлерита (1860 - 1929) для обработки данных. Для работы табулятора использовались перфокарты





• Электронные вычислительные машины появились в конце 1940-х годов. В настоящее время обработка данных переписи населения (больших данных) полностью автоматизирована, хотя в сборе данных до сих пор участвуют интервьюеры.







Где есть большие данные?

- IT
- Реклама
- Торговля
- Мобильные технологии
- Банковская сфера

- Телекоммуникации
- Энергетика
- Логистика
- Промышленность
- Метеорология

Источники больших данных









Основные характеристики BigData

- volume (объем)
- velocity (скорость)
- variety (разнообразие)

- value (ценность),
- veracity (достоверность),
- viability (жизнеспособность)
- variability (переменчивость)

Большие данные в современной жизни













Основные источники BigData

- Интернет (социальные сети, средства массовой информации, сайты, форумы и блоги)
- Корпоративные архивы
- Показания различных приборов и устройств

Чем помогает обработка больших данных?

Примеры:

- 1. Как погода влияет на продажи
- 2. Где территориально лучше распространять тот или иной продукт
- 3. Где лучше расположить парковки
- 4. Управление загрузкой горнолыжных курортов
- 5. Расположение наружной рекламы и частота ее появления

Проблема 2000 года

Компьютеры получили достаточно широкое распространение в 1960-е годы, когда до 2000 года было еще очень далеко. Разработчики программного обеспечения в XX веке часто использовали для обозначения года только две последние цифры. Соответственно, многие системы предполагали, что год начинается с «19». То есть при наступлении следующего века такие системы могли предположить, что 2015 год — это 1915 год. Подобное могло привести к серьезным сбоям в работе финансовых программ и систем управления технологическими процессами. Программы вообще могли прекратить работать в 2000 году.

Проблема возникла из-за того, что разработчики программ не подумали о том, что может произойти при смене столетий. Было приложено немало усилий, и, по некоторым данным, потрачено свыше 300 миллиардов долларов.

Техники и методы анализа, применимые к Big data по McKinsey

- Data Mining;
- Краудсорсинг;
- Смешение и интеграция данных;
- Машинное обучение;
- Искусственные нейронные сети;
- Распознавание образов;
- Прогнозная аналитика;
- Имитационное моделирование;
- Пространственный анализ;
- Статистический анализ;
- Визуализация аналитических данных.

Технологии

- . NoSQL;
- . MapReduce;
- . Hadoop;
- . R;
- . Аппаратные решения.

SmartDataUkimet

— это проект, который позволит собирать информацию со всех ГО и из различных источников, тем самым даст уникальную возможность, применяя алгоритмы искусственного интеллекта, предоставлять аналитику и делать социально-экономические прогнозы для государства. А самое главное — позволит получать информацию без посредников. То есть, самую точную, напрямую и значительно быстрее», — отметил Асет Турысов, председатель правления АО «Национальные информационные технологии».

Вакансии

Бизнесу интересны большие данные, огромный интерес проявляют банки, да и в принципе любой крупный или средний бизнес, который хранит данные. Пример тому — «Казпочта» и «Казахтелеком».

Рынок по обработке больших данных очень перспективен, если смотреть в будущее. В стране отсутствуют специалисты, спрос на которых с каждым годом увеличивается. Средняя заработная плата на рынке Data Scientist составляет от 200 000 тенге и доходит до 500 000 тенге в месяц и выше.