Терин Д.В.

Технологии хранения информации и больших объемов данных

Большие данные – разнообразные данные, которые поступают с постоянно растущей скоростью и объём котоырх постоянно растет

Три основных свойства больших данных – разнообразие, высокая скорость поступления и большой объем

Примеры:

Умные устройства

Бизнес

Здравоохранение

Создание флэш памяти – обеспечила быструю обработку данных

Обзор 2020 компании Data Age report по технологической цифровизации до 2025 года

Необходимость RTM обработки, низкая задержка, нерегламентированный характер использования и тяжесть последствий, если данные станут недоступны.

Задачи обработки больших данных

Главной задачей обработки больших данных на сегодняшний день является максимизация пользы от накопленных данных о потреблении ресурсов или услуг.

Накопленные исторические данные и оперативные данные о потреблении услуг обладают информацией о трендах, тенденциях и изменчивости вектора предпочтений пользователей

Обработка больших данных позволяет получить пользу из исторических данных в сферах бизнеса, здравоохранения, сельского хозяйства, и.т.д.

Проблемы

Проблемы технологического характера - они должны собираться независимо и незаметно от поставщика больших данных.

Проблема по созданию цифровых платформ по обработки данных

Решения

Выделяют следующих специалистов

Инженер данных

Аналитик данных

Разработчик

DEVops (в процессе используется большая группа программистов, он обеспечивает чтобы все имели одну и туже версию и чтобы использовали одни и те же библиотеки и одну и ту же версию), MLops

Управляющий продуктом

Инженер данных

Направления работы инженера данных

Предоставление данных для аналитики

Предоставление данных для машинного обучения

Внедрение моделей машинного обучения в продукт

Сбор данных из различных источников

Перемещение даннах потоки данных

Очистка подготовка трансофрмация и обработка данных по бизнес-правилам

Анализ агрегация разметка данных

Изучение данных оптимизация хранения и обработки данных

Построение платформ данных

Конвейер обработки данных

Состоит из 3 основных составляющих

Собрать

Очистить

Сохранить в нужном виде и затем работать с ними

В современной индустрии устоялся классический конвейер работы с данными, как основной шаблон по которому компания подстраиваются поток работ под себя

Под каждую новую задачу поток работ над данными видоизменяется при неизменной основе

Видоизменения набора задач по обработке данных зависит от количества источников данных, сложности данных и целей обработки данных

Инфраструктура обработки данных

Большие компании хранят, обрабатывают и анализируют данные на северных вычислительных устройствах или ЦОД разной степени доступности:

Вычислительная инфраструктура:

1 Локальный вычислительный кластер

2 Частные облачные сервисы общедоступное облако

Популярные серверные ОС

Linux – серверные системы

Debian/CentOS