Технологии работы с большими данными

Предварительный план занятий

- Разберем основной стек технологий по работе с Big Data
- SQL leetcode + Polars + повторим Pandas
- Многопоточность и многопроцессорность на Python
- Посмотрим митапы по Big Data инфре от разных компаний

Введение в Big Data

Определение: Большие данные — это наборы данных настолько больших и сложных, что традиционные инструменты не могут их обрабатывать

Характеристики больших данных: 3V (Volume, Velocity, Variety)

Объем (Volume) — огромные массивы данных.
Скорость (Velocity) — быстрая генерация данных.
Разнообразие (Variety) — данные в разных форматах (структурированные, неструктурированные, полу-структурированные).

Введение в Big Data

Источники данных

Транзакционные данные (банковские операции, покупки и т.д.)
Лог-файлы веб-сайтов и приложений
Социальные сети (Twitter, Facebook и др.)
Сенсоры и IoT-устройства
Медицинские данные

Технологии обработки больших данных (на слуху)

Hadoop — система для распределенной обработки больших данных. Основные компоненты: HDFS (распределенная файловая система) и MapReduce (метод обработки данных)
Spark — распределенная система обработки данных в реальном времени
NoSQL базы данных (Cassandra, MongoDB, HBase)
Kafka — система потоковой передачи данных

Введение в Big Data

Примеры использования больших данных

Прогнозирование потребностей клиентов (ретейл, финансы)
Обнаружение мошенничества (банковская сфера)
Анализ поведения пользователей в интернете
IoT и умные города (умные счетчики, транспортные системы)

Тренды в области больших данных

Искусственный интеллект и машинное обучение для анализа данных
Потоковая обработка данных в реальном времени
Edge Computing (вычисления на уровне устройств)
Развитие облачных технологий для масштабирования хранения и обработки данных

Почему Big Data важна?

Применение:
Аналитика в бизнесе: предсказание потребностей клиентов.
Обработка данных в реальном времени: обработка логов, мониторинг систем.
Научные исследования: анализ данных геномики, космические исследования.

Ценность:
Big Data помогают принимать более обоснованные решения и автоматизировать процессы.

Основные этапы работы с большими данными

Сбор данных: использование сенсоров, API, веб-скрапинга.
Хранение данных: распределенные системы (HDFS, облачные хранилища)
Предобработка данных: очистка, нормализация, агрегация данных
Анализ данных: использование инструментов машинного обучения и статистики
Визуализация данных: создание графиков и диаграмм для представления результатов

Что студентам нужно учить?

Основы программирования:
Python, SQL — для обработки и анализа данных.
Простые задачи на чтение, запись и фильтрацию данных.

Базы данных:
Реляционные базы данных (MySQL, PostgreSQL).
NoSQL базы данных (MongoDB, Cassandra) — для работы с неструктурированными данными.

Технологии Big Data:
Hadoop: установка кластера, работа с HDFS и MapReduce.
Spark: запуск простых задач и работа с PySpark.

Аналитика и машинное обучение:
Основы статистики и машинного обучения (Scikit-learn, MLlib в Spark).

Полезные материалы

Крутой курс от VK (обязательный к просмотру)
Машинное обучение для больших данных
Много обучающих материалов