Наименование

Курс "Алгоритмы и технологии обработки больших данных",

Продолжительность

1 год, лекции по 1 часу в неделю, семинарские занятия по 1 часу в неделю

Преподаватели

Лектор: Марчук А.Г., д.ф.-м.н., профессор

Ассистент: Платонов Ю.Г., к.ф.-м.н.

Структура курса

Структуризация данных, базы данных, СУБД, реляционные базы данных и СУБД.

Введение в язык программирования C#

Типы данных, типизованные структуры, отображение структур на среды: текст, ОЗУ, потоки байтов, библиотека PolarDB.

Стандартные представления структурированных данных, XML, реляционные таблицы, JSON, RDF.

Потоки и их обработка, функциональное программирование средствами Linq.

Последовательности, как основа работы со структурированными данными. Индексы и их построение.

NoSQL СУБД, примеры: MongoDB, Cassandra.

Основы клиент-сервисной архитектуры, протоколы, стандартные протоколы, протокол TCP.

Построение распределенных систем хранения/обработки, облачные вычисления сервисные инфраструктуры, примеры: Amazon, MS Azure.

Структурирование графовых данных, связь с реляционными построениями, RDF, атрибутированные графы, Sparql

Открытые архитектуры сетевой обработки, Hadoop, Dynamo, MapReduce и др.

Спецификации структур данных, онтологии, формализм OWL, сравнение онтологий, связанные данные (LOD).

Элементы обработки текстов, поиск по ключевым словам, использование словарей, n-Граммы.

Целевая аудитория

Магистранты, желающие существенно углубить знания в области технологий современной работы с данными.

Курс может быть полезен также учащимся в бакалавриате, аспирантам, молодым специалистам, продвинутым школьникам.

Техническое оснащение

Предполагается работа в компьютерной аудитории так, чтобы студенты могли выполнять задания.

--А.Г.

# Алгоритмы и технологии обработки больших данных

Учебный курс, 1 год (1 ч лекция и 1 ч семинара в неделю), 20 недель

## Введение

Активная компьютеризация практически всех областей жизни и производственной деятельности, вводит в работу очень большие объемы структурированной, полуструктурированной и неструктурированной информации. Этому способствует постоянный и быстрый рост объемов накопителей информации, удешевление их стоимости. В итоге, есть потребность и техническая возможность работы с большими и очень большими объемами данных. Но не всегда есть способы такой работы. Это касается и методов обработки и алгоритмов. Как правило, с ростом объема данных, известные решения перестают «работать», либо их работа становится слишком медленной. Все это способствовало формированию нового направления фундаментальных и прикладных исследований – Большие данные (Big Data). На изучение основных аспектов данной темы направлен данный курс.

Курс является практико-ориентированным, т.е. кроме теории предмета, дается возможность достаточно глубоко изучить и алгоритмы, лежащие в основе систем обработки больших данных и технологии. Причем как имеющиеся технологии, уже хорошо зарекомендовавшие себя, так и новые, только появляющиеся. Основным языком программирования, используемым в курсе, будет C#, основной средой программирования - .NET. Многие примеры и упражнения будут выполнены или выполняться с использованием специального слоя функционального программирования Linq, хорошо интегрированного с языком C# и эффективно погруженного в библиотечные решения .NET. Предполагается, что слушатель курса или уже знает основы C# или изучит их самостоятельно.

В качестве инструментальной платформы программирования рекомендуется Visual Studio. Для всех задач достаточно использования официальной бесплатной версии студии, ориентированной на обучение и некоммерческое программирование. Полезно также будет установить и использовать в некоторых случаях автономную среду программирования .Net Core, ориентированную на работу с консолью. К такому решению удобно добавить универсальный редактор кода Visual Code, который не только существенно «легче» студии, но и реактивнее и нагляднее.

## Hello World!

Традиционным началом

## Структуризация данных