МИНОБРНАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ И КОМПЬЮТЕРНЫХ НАУК

УТВЕРЖДАЮ
Директор института прикладной
математики и компьютерных наук
_____ А.В. Замятин
«_14_» _____ 2021 г.

Анализ больших массивов данных (Big Data)

Рабочая программа дисциплины

Закреплена за кафедрой	*теоретических основ информатики*
Учебный план	*01.03.02 Прикладная математика и информатика, профиль «Прикладная математика и информатика»*
Форма обучения	*очная*
Общая трудоёмкость	*2 з.е.*
Часов по учебному плану	*72*
в том числе:	
аудиторная контактная работа	*33,85*
самостоятельная работа	*38,15*
Вид(ы) контроля в семестрах *экзамен/зачёт/зачёт с оценкой*	*Семестр 8 – зачёт*

Томск–2021

Программу составил:
д.т.н., профессор,
зав. кафедрой теоретических основ информатики А.В. Замятин

Рецензент:
д.т.н., профессор,
заведующий кафедрой прикладной информатики С.П. Сущенко

Рабочая программа дисциплины «Анализ больших массивов данных (Big Data)» разработана в соответствии с самостоятельно устанавливаемым образовательным стандартом высшего образования – бакалавриат – Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский Томский государственный университет» по направлению подготовки 01.04.02 Прикладная математика и информатика (Утверждён Учным советом НИ ТГУ, протокол от 27.10.2021 г. № 08).

Рабочая программа одобрена на заседании кафедры теоретических основ информатики

Протокол от 04 июня 2021 г. № 05

Заведующий кафедрой теоретических основ информатики,
д-р техн. наук, профессор А.В. Замятин

Рабочая программа одобрена на заседании учебно-методической комиссии института прикладной математики и компьютерных наук (УМК ИПМКН)

Протокол от 17.06.2021 г. № 05

Председатель УМК ИПМКН,
д.т.н., профессор С.П. Сущенко

Цель и задачи дисциплины

Знакомство с современными методами и инструментами обработки больших массивов данных.

Задачами дисциплины являются:

- изучение математических и алгоритмических методов обработки больших массивов данных;
- знакомство со стеком технологий обработки больших массивов данных Hadoop;
- приобретение практического опыта применения инструментов обработки больших массивов данных для решения практических задач.

1. Место дисциплины в структуре ОПОП

Дисциплина «Анализ больших массивов данных (Big Data)» относится к обязательной части профессионального цикла блока Б1.П.Блок Б1. Дисциплины (модули).

Пререквизиты дисциплины: нет

Постреквизиты дисциплины: нет

2. Компетенции и результаты обучения, формируемые в результате освоения дисциплины

Компетенция	Индикатор общепрофессиональной компетенции	Код и наименование результатов обучения (планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций)
ОПК-2 Способен использовать и адаптировать существующие математические методы и системы программирования для разработки и реализации алгоритмов решения прикладных задач	ИОПК-2.1 Обладает навыками объектно-ориентированного программирования для решения прикладных задач в профессиональной деятельности	
	ИОПК-2.2 Проявляет навыки использования основных языков программирования, основных методов разработки программ, стандартов оформления программной документации	
	ИОПК-2.3 Демонстрирует умение отбора среди существующих математических методов, наиболее подходящих для решения конкретной прикладной задачи	
	ИОПК-2.4 Демонстрирует умение адаптировать существующие математические методы для решения конкретной прикладной задачи	
ОПК-4 Способен решать задачи профессиональной деятельности с использованием	ИОПК-4.1 Проявляет владение базовыми знаниями по защите информации на рабочем месте и при входе в локальные и глобальные сети	

существующих информационно-коммуникационных технологий и с учетом основных требований информационной безопасности	ИОПК-4.2 Демонстрирует навыки использования научных и образовательных ресурсов сети Интернет для разработки программ и программной документации с учетом требований информационной безопасности	
ПК-3 Способен формализовывать, согласовывать и документировать требования к системе и подсистеме, обрабатывать запросы на изменение требований к системе и подсистеме, выявлять и формализовывать риски, анализировать проблемные ситуации	ИПК 3.1 Реализовывает построение формализованной математической модели системы (подсистемы): введение целевой функции системы (подсистемы) и ограничений, соответствующих требованиям к системе (подсистеме)	
	ИПК 3.2 Адаптирует формализованную математическую модель системы (подсистемы) к изменению требований (ограничений к целевой функции) к системе (подсистеме)	
	ИПК 3.3 Выявляет и формализовывает в виде математической модели возникающие при функционировании системы (подсистемы) риски; выявляет и анализирует проблемные ситуации	

3. Структура и содержание дисциплины

3.1 Структура и трудоёмкость видов учебной работы по дисциплине

Общая трудоёмкость дисциплины составляет 3 зачётных единицы, 108 часов.

Вид учебной работы	Трудоёмкость в академических часах	
	8 семестр	всего
Общая трудоёмкость	72	72
Контактная работа:	33,85	33,85
Лекции (Л)	16	16
Практики (ПЗ)		
Лабораторные работы (ЛР)	16	16
Семинары (СЗ)		
Групповые консультации		
Индивидуальные консультации	1,6	1,6
Промежуточная аттестация	0,25	0,25
Самостоятельная работа обучающегося:	38,15	38,15
- выполнение контрольных заданий	8	8
- изучение учебного материала	14	14
- подготовка к практическим занятиям/коллоквиумам	14	14
- подготовка к рубежному контролю по теме/разделу	2,15	2,15
Вид промежуточной аттестации (зачет, зачет с оценкой, экзамен)	**Зачет**	

3.2 Содержание и трудоёмкость разделов дисциплин

Код занятия	Наименование разделов и тем и их содержание	Вид учебной работы, занятий, контроля	Семестр	Часы в электронной форме	Всего (час.)	Литература	Код(ы) результата(ов) обучения
Раздел 1. Hadoop							
1.1	Экосистема Hadoop. Файловая система HDFS. Инструменты хранения и обработки данных Flume, MySQL, Hive.	Лекция	3	-	4	№1, №2, №5	
1.2		Лаб.работа	3	-	4		
1.3		СРС	3	-	7		
1.4		Контр.работа	3	-	2		
Раздел 2. Язык программирования Python							
2.1	Основы программирования на языке Python. Библиотеки анализа данных numpy, pandas, matplolib.	Лекция	3	-	4	№4	
2.2		Лаб.работа	3	-	4		
2.3		СРС	3	-	7		
2.4		Контр.работа	3	-	2		
Раздел 3. Модель вычислений Map/Reduce							
3.1	Модель распределённых вычислений Map/Reduce и шаблоны решения типовых задач анализа данных	Лекция	3	-	4	№1, №2, №4	
3.2		Лаб.работа	3	-	4		
3.3		СРС	3	-	8		
3.4		Контр.работа	3	-	2		
Раздел 4. Платформа распределённой обработки Spark							
4.1	Модель распределённых вычислений на платформе Spark. Решение типовых задач анализа данных с применением расширений Spark SQL, Spark MLib	Лекция	3	-	4	№3, №4, №6, №7	
4.2		Лаб.работа	3	-	4		
4.3		СРС	3	-	8,15		
4.4		Контр.работа	3	-	2		

Содержание тем дисциплины

Тема 1. История возникновения и развития системы распределённых вычислений Hadoop. Компоненты системы Hadoop и их взаимозависимости. Распределённая файловая система HDFS (архитектура, свойства и основные команды). Инструменты хранения и обработки данных Flume, MySQL, Hive.

Тема 2. Основные понятия зыка Python (типы и структуры данных, объектно-ориентированное программирование, работа с файлами). Библиотека математических вычислений numpy. Библиотека анализа данных pandas. Библиотека визуализации данных matplotlib.

Тема 3.Модель вычислений Map/Reduce. Реализация модели вычислений Map/Reduce в системе Hadoop. Решение типовых задач анализа данных (основные теоретико-множественные операции и операции реляционной алгебры) в рамках модели Map/Reduce.

Тема 4. Архитектура платформы распределённой обработки Spark. Основные типы данных Spark и типы операций Spark, выполняемых на узлах кластера. Решение типовых задач анализа данных с применением базовых типов данных Spark и его расширений Spark SQL и Spark MLib.

4. Образовательные технологии, учебно-методическое и информационное обеспечение для освоения дисциплины

Основной технологией освоения дисциплины являются лекции, материал которых закрепляется путём решения практических задач на лабораторных практикумах. Для контроля усвоения материала по каждой теме предусмотрена контрольная работа.

Промежуточная аттестация осуществляется в виде письменного тестирования знания прослушанного материала при условии успешного решения всех контрольных работ.

4.1 Рекомендуемая литература

№ п/п	Авторы/составители	Заглавие	Издательство	Год издания
Основная литература				
1	Уайт Т.	Hadoop. Подробное руководство	СПб.: Питер	2013
2	Чак Лэм	Hadoop в действии	М.: ДМК Прес	2012
3	Карау Х., Конвински Э., Венделл П., Захария М.	Изучаем Spark: молниеносный анализ данных	М.: ДМК Пресс	2015
4	Лутц М.	Изучаем Python	СПб.: Питер	2011
5	Компания MySQL AB	MySQL. Справочник по языку	М. Вильямс	2005
Дополнительная литература				
6	Карау Х., Уоррен Р.	Эффективный Spark. Масштабирование и оптимизация	СПб.: Питер	2018
7	Риза С., Лезерсон У., Оуэн Ш., Уилисс Дж.	Spark для профессионалов. Современные паттерны обработки больших данных	СПб.: Питер	2017

4.2 Базы данных и информационно-справочные системы, в том числе зарубежные

- Электронный каталог Научной библиотеки ТГУ http://www.lib.tsu.ru/
- Документация по языку Python https://docs.python.org/3/contents.html
- Официальный сайт проекта NumPy http://www.numpy.org/
- Официальный сайт проекта Pandas http://pandas.pydata.org/
- Официальный сайт проекта Matplotlib https://matplotlib.org/
- Официальный сайт документации проекта MySQL https://dev.mysql.com/doc/
- Официальный сайт проекта Hadoop http://hadoop.apache.org/
- Официальный сайт проекта Sqoop http://sqoop.apache.org/
- Официальный сайт проекта Flume http://flume.apache.org/
- Официальный сайт проекта Hive https://hive.apache.org/
- Официальный сайт проекта Spark https://spark.apache.org/

4.3 Перечень лицензионного и программного обеспечения

- Операционная система – 64 разрядная Windows (версии 7, 8 или 10);
- Система виртуализации Oracle Virtual Box;
- Интерпретатор Python версии 3 (дистрибутив Anaconda);
- Пакет Microsoft Office (Word, Excel, Access, PowerPoint).

4.4 Оборудование и технические средства обучения

К аудитории для занятий предъявляются следующие требования:

- На рабочих местах установлены персональные компьютеры со следующей конфигурацией:

- центральный процессор – Intel Core i7 (или аналогичный),
- объёмом оперативной памяти – не менее 8Gb,
- свободное место на жестком диске – не менее 100Gb,
- подключение к локальной сети – не менее 100Mb/s (желательно 1000 Gb/s).

- Аудитория оборудована маркерной доской и проектором.

5. Методические указания обучающимся по освоению дисциплины

Основой обучения являются материалы лекций и лабораторных занятий. Для самостоятельного изучения и расширения круга знаний рекомендуется использовать предлагаемую литературу и актуальную документацию по проектам в сети Интернет.

6. Преподавательский состав, реализующий дисциплину

Богданов Александр Леонидович, к.т.н., доцент, доцент кафедры информационных технологий и бизнес-аналитики ИЭМ ТГУ.

7. Язык преподавания – русский.