МИНПРОСВЕЩЕНИЯ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования "Тульский государственный педагогический университет им. Л.Н. Толстого"

(ФГБОУ ВО "ТГПУ им. Л.Н. Толстого")

Технологии и методы обработки больших данных

рабочая программа дисциплины (модуля)

Закреплена за кафедрой **институт передовых информационных технологий**

ОПОП **Направление 02.03.02 Фундаментальная информатика и информационные технологии**
направленность (профиль) Большие данные и распределенная цифровая платформа

Квалификация **Бакалавр**

Год начала подготовки **2023**

Форма обучения **очная**

Общая трудоемкость **4 з.е.**

Виды контроля по семестрам:
зачет 5

Семестр(Курс.Номер семестра на курсе)	**5(3.1)**		Итого	
	УП	РПД	УП	РПД
Лекции	30	30	30	30
Лабораторные	30	30	30	30
Итого ауд.	60	60	60	60
Контактная работа	62	62	62	62
Сам. работа	80	80	80	80
Часы на контроль	2	2	2	2
Практическая подготовка	0	0	0	0
Семинары	0	0	0	0
Консультации	2	2	2	2
Итого трудоемкость в часах	146	146	146	146

Программу составил(и):

д.ф.-м.н., доцент, Ганкевич Иван Геннадьевич

Рабочая программа дисциплины

Технологии и методы обработки больших данных

разработана в соответствии с ФГОС:

Федеральный государственный образовательный стандарт высшего образования - бакалавриат по направлению подготовки 02.03.02 Фундаментальная информатика и информационные технологии (приказ Минобрнауки России от 23.08.2017 г. № 808)

составлена на основании учебного плана:

Направление 02.03.02 Фундаментальная информатика и информационные технологии
направленность (профиль) Большие данные и распределенная цифровая платформа
утвержденного Учёным советом вуза от 29.09.2022 протокол № 11.

РПД утверждена Учёным советом университета
протокол от 29.9.2022 г. № 11

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Целью дисциплины является изучение различных аспектов технологий хранения и обработки больших объемов данных, таких как автоматический сбор и консолидация, пакетная и потоковая обработка, обработка в реальном времени, анализ текстов на естественном языке.

В задачу курса входит обучение студентов

- работе с параллельными и распределенными файловыми системами,
- разработке программ для параллельной пакетной и потоковой обработки больших объемов данных и
- особенностям реализации некоторых алгоритмов для больших объемов данных.

По окончании курса студент приобретет навыки и умения, которые широко применяются в отрасли информационных технологий для анализа эффективности взаимодействия поставщика и потребителя.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП

Цикл (раздел) ООП:	Б1.В.ДВ.01

2.1	**Требования к предварительной подготовке обучающегося:**
1.	Технологии искусственного интеллекта
2.	Основы распределенных вычислений
3.	Системное программирование в Linux
4.	Современные методы программирования
5.	Функциональное программирование
6.	Архитектура вычислительных систем
2.2	**Дисциплины и практики, для которых освоение данной дисциплины (модуля) необходимо как предшествующее:**
1.	Верификация, аттестация и качество программного обеспечения
2.	Вычисления общего назначения на видеокарте
3.	Криптография и блокчейн
4.	Машинное обучение
5.	Методы и средства научной визуализации
6.	Учебная практика (научно-исследовательская работа)
7.	Нейросетевые технологии
8.	Облачные и высокопроизводительные вычисления
9.	Проектирование баз данных для сложных информационных систем
10.	Вариационные задачи обработки изображений
11.	Основы научной коммуникации
12.	Производственная практика (научно-исследовательская работа) (на английском языке)

3. СООТНЕСЕНИЕ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ (МОДУЛЮ) С ИНДИКАТОРАМИ ДОСТИЖЕНИЯ КОМПЕТЕНЦИЙ

3.1 Компетенции обучающегося и индикаторы их достижения:

ОПК-2: Способен применять компьютерные/суперкомпьютерные методы, современное программное обеспечение, в том числе отечественного происхождения, для решения задач профессиональной деятельности	
ОПК-2.1	Знает современные информационные технологии и программные средства, в том числе отечественного производства при решении задач профессиональной деятельности
	принципы построения систем обработки и анализа больших объемов данных
ОПК-2.2	Умеет выбирать современные информационные технологии и программные средства, в том числе отечественного производства при решении задач профессиональной деятельности
	выбрать подходящую информационную систему для наиболее эффективного конкретной решаемой задачи
ОПК-2.3	Владеет навыками применения современных информационных технологий и программных средств, в том числе отечественного производства, при решении задач профессиональной деятельности
	навыками разработки программ пакетной обработки данных с помощью Hadoop
ОПК-4: Способен участвовать в разработке технической документации программных продуктов и комплексов с использованием стандартов, норм и правил, а также в управлении проектами создания информационных систем на стадиях жизненного цикла	
ОПК-4.1	Знает принципы сбора и анализа информации, создания информационных систем на стадиях жизненного цикла
	базовые знания математических и естественных наук, программирования и информационных технологий
ОПК-4.2	Умеет осуществлять управление проектами информационных систем

	использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков программирования и пакетов прикладных программ моделирования
ОПК-4.3	Имеет практический опыт анализа и проектирования информационных систем
	навыками разработки программ пакетной обработки данных с помощью Spark
ПК-1: Способен профессионально заниматься разработкой и внедрением новых технологий цифровой экономики	
ПК-1.1	Оценивает возможности применения различных архитектур вычислительных систем для решения различных задач цифровой экономики
	владеть навыками работы с параллельными и распределенными файловыми системами (таких как GlusterFS и HDFS)
ПК-1.2	Способен планировать состав вычислительных средств для решения поставленных задач
	Способен использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков программирования и пакетов прикладных программ моделирования
ПК-2: Способен обрабатывать и структурировать разнородную статистическую экспериментальную и экономико-производственную информацию с использованием территориально-распределенных технологических ресурсов и цифровых платформ	
ПК-2.1	Знает современные приемы работы с инструментальными средствами, поддерживающими создание программных продуктов и программных комплексов, их сопровождения и администрирования
	базовые знания математических и естественных наук, программирования и информационных технологий
ПК-2.2	Умеет использовать подобные инструментальные средства в практической деятельности
	использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков программирования и пакетов прикладных программ моделирования
ПК-2.3	Владеет практический опыт применения подобных инструментальных средств
	Способен творчески применять базовые знания математических и естественных наук, программирования и информационных технологий
ПК-3: Способен реализовывать концепции развития и использования технологий Больших данных и высокопроизводительных вычислений в рамках структур академической науки, экономической деятельности и государственного управления	
ПК-3.1	Знает направления развития компьютеров с традиционной (нетрадиционной) архитектурой; современных системных программных средств; операционных систем, операционных и сетевых оболочек, сервисных программ; тенденции развития функций и архитектур проблемно ориентированных программных систем и комплексов в профессиональной деятельности
	принципы построения систем обработки и анализа больших объемов данных
ПК-3.2	Умеет программировать для компьютеров с различной современной архитектурой
	владеть навыками разработки программ пакетной обработки данных с помощью Hadoop
ПК-3.3	Владеет практический опыт выбора архитектуры и комплексирования современных компьютеров, систем, комплексов и сетей системного администрирования
	навыками разработки программ пакетной обработки данных с помощью Spark

3.2 Результаты обучения по дисциплине:
В результате освоения дисциплины обучающийся должен:

	Знать:
З.1	базовые знания математических и естественных наук, программирования и информационных технологий
З.2	принципы построения систем обработки и анализа больших объемов данных
	Уметь:
У.1	использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков программирования и пакетов прикладных программ моделирования
У.2	выбрать подходящую информационную систему для наиболее эффективного конкретной решаемой задачи
	Владеть:
В.1	навыками разработки программ пакетной обработки данных с помощью Hadoop
В.2	навыками разработки программ пакетной обработки данных с помощью Spark
В.3	владеть навыками работы с параллельными и распределенными файловыми системами (таких как GlusterFS и HDFS)

4. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Литература	Содержание
	Пакетная обработка больших данных в Hadoop				

1.1	Понятие больших данных. Свойства больших данных, свойства 5V. Алгоритм MapReduce, краткая история и предпосылки появления. Экосистема Hadoop, YARN, HDFS. Понятия репликации, отказоустойчивости. Основные классы: Mapper, Reducer, Job, *Writable. Запуск нескольких задач одновременно. /Лек/	5	4	Л1.1Л2.1	
1.2	Понятие больших данных. Свойства больших данных, свойства 5V. Алгоритм MapReduce, краткая история и предпосылки появления. Экосистема Hadoop, YARN, HDFS. Понятия репликации, отказоустойчивости. Основные классы: Mapper, Reducer, Job, *Writable. Запуск нескольких задач одновременно. /Лаб/	5	6	Л1.1Л2.1	
1.3	Понятие больших данных. Свойства больших данных, свойства 5V. Алгоритм MapReduce, краткая история и предпосылки появления. Экосистема Hadoop, YARN, HDFS. Понятия репликации, отказоустойчивости. Основные классы: Mapper, Reducer, Job, *Writable. Запуск нескольких задач одновременно. /Ср/	5	20	Л1.1Л2.1	
	Основы анализа текстов с помощью Lucene				
2.1	Библиотеки Lucene, Tika, веб-сервис Solr. Реализация разбиения на ключевые слова, индексации, поиска, написание запросов. Параллельная индексация больших объемов данных с помощью Hadoop. /Лек/	5	8	Л1.1Л2.1	
2.2	Библиотеки Lucene, Tika, веб-сервис Solr. Реализация разбиения на ключевые слова, индексации, поиска, написание запросов. Параллельная индексация больших объемов данных с помощью Hadoop. /Лаб/	5	8	Л1.1Л2.1	

2.3	Библиотеки Lucene, Tika, веб-сервис Solr. Реализация разбиения на ключевые слова, индексации, поиска, написание запросов. Параллельная индексация больших объемов данных с помощью Hadoop. /Ср/	5	20	Л1.1Л2.1	
	Анализ больших данных в Spark				
3.1	Предпосылки создания. Преимущества и недостатки по сравнению с Hadoop. Понятие отказоустойчивых распределенных массивов, преобразований и действий. Программирование с использованием функциональных примитивов и автоматическое построение графа задач. /Лек/	5	8	Л1.1Л2.1	
3.2	Предпосылки создания. Преимущества и недостатки по сравнению с Hadoop. Понятие отказоустойчивых распределенных массивов, преобразований и действий. Программирование с использованием функциональных примитивов и автоматическое построение графа задач. /Лаб/	5	8	Л1.1Л2.1	
3.3	Предпосылки создания. Преимущества и недостатки по сравнению с Hadoop. Понятие отказоустойчивых распределенных массивов, преобразований и действий. Программирование с использованием функциональных примитивов и автоматическое построение графа задач. /Ср/	5	20	Л1.1Л2.1	
	Потоковая обработка данных в Spark				
4.1	Особенности программной реализации потоковой обработки в Spark. Отличие потоковой обработки от обработки в реальном времени. /Лек/	5	10	Л1.1Л2.1	

4.2	Особенности программной реализации потоковой обработки в Spark. Отличие потоковой обработки от обработки в реальном времени. /Лаб/	5	8	Л1.1Л2.1	
4.3	Особенности программной реализации потоковой обработки в Spark. Отличие потоковой обработки от обработки в реальном времени. /Ср/	5	20	Л1.1Л2.1	
	Консультация				
5.1	Консультация /Конс/	5	2	Л1.1Л2.1	

5. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ

5.1. Типовые задания для проведения текущего контроля

Обработка текстов на русском языке с помощью библиотеки Lucene. Использование классов из предметной области задачи в Hadoop. Сортировка выходных данных.
Предварительная обработка частотно-направленных спектров морского волнения. Очистка данных. Работа со структурированными данными особого формата.
Преобразование данных из текстового формата в формат NetCDF.
Обработка текстов на русском языке. Переписывание программы для обработки спектров с Hadoop на Spark.
Обработка потока данных из системного журнала.

5.2. Типовые задания для проведения промежуточной аттестации

Примеры практических заданий.
1. Перепишите программу, подсчитывающую частоту использования слов в тексте, используя библиотеку Lucene: учтите все возможные разделители слов и исключите стоп-слова из вывода. Для этого воспользуйтесь стандартным анализатором.
2. Измените программу из задания 1, так чтобы посчитать количество слов определенной длины в тексте. Результатом работы программы должна стать таблица вида {длина слова → количество слов такой длины}. Ключ должен иметь интегральный тип.
3. Измените программу из задания 1, так чтобы для каждого слова вывести слово, чаще всего следующее за ним. Учтите знаки препинания, являющиеся разделителями для предложений, чтобы исключить из вывода слова, идущие друг за другом, но находящиеся в разных предложениях.
4. Измените программу из предыдущего задания так, чтобы дополнить вывод частотой встречи слов (количество раз, которое слово-значение следует за словом-ключом). Для этого создайте новый выходной тип с соответствующими полями, реализовав интерфейс org.apache.hadoop.io.Writable.
5. Измените программу из предыдущего задания так, чтобы упорядочить вывод по убыванию частоты встречи слов. Для этого необходимо реализовать интерфейс org.apache.hadoop.io.WritableComparable у созданного вами типа и дать планировщику дополнительное задание, которое отсортирует данные.

5.3. Перечень видов оценочных средств

Собеседование по программе курса в части пройденного материала, отчётность по выполнению самостоятельных работ. Подготовка формальных отчетов. Знание основных определений объектов, упоминаемых в программе, выполненная практическая работа по курсу, продемонстрированные результаты практической работы по курсу..
Зачет.

5.4. Процедура применения оценочных материалов

Оценка ECTS выставляется на основе итогового процента выполнения итогового теста

Итоговый процент выполнения, %	Оценка СПбГУ при проведении зачёта	Оценка ECTS
90-100	зачтено	A
80-89	зачтено	B
70-79	зачтено	C
61-69	зачтено	D
50-60	зачтено	E
менее 50	не зачтено	F

6. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

6.1. Рекомендуемая литература

6.1.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год (кол-во экземпляров для печатных изданий)	Ссылка на электронное издание

	Авторы, составители	Заглавие	Издательство, год (кол-во экземпляров для печатных изданий)	Ссылка на электронное издание
Л1.1	Мендель А. В.	Модели принятия решений: учебное пособие	Москва : Юнити-Дана, 2015	http://biblioclub.ru/index.php?page=book&id=115173
6.1.2. Дополнительная литература				
	Авторы, составители	Заглавие	Издательство, год (кол-во экземпляров для печатных изданий)	Ссылка на электронное издание
Л2.1	под ред. В. Г. Халина, Г. В. Черновой	Системы поддержки принятия решений: учебник и практикум для бакалавриата и магистратуры	Москва: Издательство Юрайт, 2016	http://www.biblio-online.ru/book/8D604E99-FC0E-4483-9F5E-54AAD6B89852

6.2. Перечень ресурсов информационно-телекоммуникационной сети "Интернет"	
Э1	Сайт Научной библиотеки им. М. Горького СПбГУ
Э2	Перечень электронных ресурсов, находящихся в доступе СПбГУ
Э3	Электронный каталог Научной библиотеки им. М. Горького СПбГУ
Э4	Перечень ЭБС, на платформах которых представлены российские учебники, находящиеся в доступе СПбГУ
6.3.Информационные технологии	
6.3.1 Перечень лицензионного и свободно распространяемого программного обеспечения	
1.	Операционная система ROSA Enterprise Linux Desktop № RL00450-1-110518-01. RL00450-1-110518-17 от 11 мая 2018 г.
2.	Операционная система Microsoft Windows XP Professional Russian. Лицензия № 16698685 от 08.08.2003 г.
3.	Операционная система Microsoft Windows Professional 7 Russian. Лицензия №48497058 от 13.05.2011 г., договор № Пр/16/6 от 05 апреля 2016 г.
4.	Операционная система Microsoft Windows 10 Professional Russian. Контракт № ПР/ФЕН/15/18 от 23.10.2015 г., договор № Пр/16/6 от 05 апреля 2016 г.
5.	Программное обеспечение Microsoft Office Enterprise 2007 Russian. Лицензия №46138962 от 16.11.2009
6.	Программное обеспечение Microsoft Office 2013 Professional. Контракт № 405535 от 2 ноября 2015 года, контракт № ПР/ФЕН/15/18 от 23.10.2015 г.
7.	Программа для распознавания текста ABBYY FineReader 9.0 Corporate Edition. Лицензионный сертификат - код позиции AF90-3U1V25-102, ABBYY FineReader 9.0 Corporate Edition Volume License Concurrent от 28 июля 2009 г.
8.	Электронный словарь ABBYY Lingvo X3 Европейская версия - Код позиции AL14-2U1V05-102, ABBYY Lingvo x3 Европейская версия. Именная лицензия Concurrent от 28 июля 2009 г.
9.	Комплексная система антивирусной защиты Kaspersky Endpoint Security для бизнеса – стандартный Russian Edition. 500-999 Node 2 year Educational Renewal License. Лицензия № 13C8-190514-084943-783-1256 от 15.05.2019
10.	Файловый архиватор 7z. Свободно распространяемое ПО
11.	Браузеры Google Chrome, Mozilla, Opera. Свободно распространяемое ПО
12.	Пакет офисных приложений Apache OpenOffice 4.1.6. Свободно распространяемое ПО
13.	Программа просмотра файлов формата RPD Adobe Acrobat Reader DC. Свободно распространяемое ПО
14.	Система Интернет-телефонии Skype. Свободно распространяемое ПО
15.	Система облачного хранилища Dropbox. Свободно распространяемое ПО
6.3.2 Перечень информационных справочных систем и профессиональных баз данных	
1.	Национальная энциклопедическая служба (https://vocabulary.ru)
2.	Базы данных издательства Springer (https://link.springer.com)
3.	Полнотекстовый архив ведущих западных научных журналов на российской платформе Национального электронно-информационного консорциума (НЭИКОН)(http://neicon.ru)
4.	Web of Science Core Collection – политематическая реферативно-библиографическая и наукометрическая (библиометрическая) база данных (http://webofscience.com)
5.	Портал «Информационно-коммуникационные технологии в образовании» (http://www.ict.edu.ru)
6.	Портал Федеральных государственных образовательных стандартов высшего образования (http://fgosvo.ru)
7.	Официальный интернет-портал базы данных правовой информации (http://pravo.gov.ru)
8.	Компьютерная информационно-правовая система «Гарант»

7. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Ауд.	Назначение	Оборудование и технические средства обучения	Вид
4-303	Помещение для самостоятельной работы	аудиоколонки, кондиционер, маркерная доска, столы компьютерные, столы учебные, компьютерная техника с возможностью подключения сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду университета	
4-305	Компьютерный класс	аудиоколонки для проектора и интерактивной доски, аудиоколонки учебные, интерактивная доска, компьютеры, кондиционер, маркерная доска, проектор, столы компьютерные, столы учебные	
4-306	Компьютерный класс	аудиоколонки для проектора и интерактивной доски, интерактивная доска, компьютеры, кондиционер, маркерная доска, проектор, столы компьютерные, столы учебные	
4-307	Компьютерный класс	аудиоколонки, компьютеры, кондиционер, маркерная доска, столы компьютерные, столы учебные, телевизор	
4-318	Компьютерный класс	компьютеры, маркерная доска, серверная стойка лаборатории МТС, стол преподавателя, столы компьютерные, столы учебный большой	
2-15	Компьютерный класс	компьютеры, рулонный экран, стол преподавателя, столы компьютерные, переносной проектор	
2-16	Компьютерный класс	интерактивная доска, компьютеры, маркерная доска, принтер, сканер, стол преподавателя, столы учебные	

8. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

Методические рекомендации преподавателю

Чтение лекций сопровождается показом презентации на проекторе или мультимедийном экране. Для более эффективного усвоения материала в презентации присутствует покадровая (послайдовая) анимация или видеоролик, схематично демонстрирующий принцип работы параллельного алгоритма. На практических занятиях демонстрируется работа с командной строкой в интерактивном режиме, сопровождаемая комментариями каждого из действий и вывода каждой из команд.

Методические указания студентам

Самостоятельная работа студента включает в себя написание программ с использованием Hadoop и Spark, изучением технической документации классов и интерфейсов этих фреймворков, а также поиском решений возникших проблем на информационных ресурсах сети Интернет.

Каждое из заданий курса сопровождается предисловием, объясняющим основные классы и методы, которые будут полезны в выполнении задания. После изучения предисловия следует обратиться к технической документации соответствующих классов и методов.