Министерство науки и высшего образования Российской Федерации

Муромский институт (филиал)

федерального государственного бюджетного образовательного учреждения высшего образования

«Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»

(МИ ВлГУ)

Кафедра *ПИн*

«УТВЕРЖДАЮ»
Заместитель директора по УР
_____Д.Е. Андрианов
21.05.2024

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Системы обработки больших данных

Направление подготовки *09.04.04 Программная инженерия*

Профиль подготовки *Технологии разработки интеллектуальных систем*

Семестр	Трудоемкость, час./зач. ед.	Лекции, час.	Практические занятия, час.	Лабораторные работы, час.	Консультация, час.	Контроль, час.	Всего (контактная работа), час.	СРС, час.	Форма промежуточного контроля (экз., зач., зач. с оц.)
1	**126 / 3,5**	**12**		**24**	**1,2**	**2,25**	**39,45**	**86,55**	**Зач.**
2	**126 / 3,5**	**20**	**14**	**36**	**4**	**2,35**	**76,35**	**23**	**Экз.(26,65)**
Итого	**252 / 7**	**32**	**14**	**60**	**5,2**	**4,6**	**115,8**	**109,55**	**26,65**

Муром, 2024 г.

1. Цель освоения дисциплины

Цели:

Осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач на основе применения больших данных;

Осуществлять сбор, обработку и статистический анализ данных, необходимых для решения поставленных экономических задач с применением методов анализа и обработки больших данных;

Понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности с использованием технологий больших данных.

Задачи:

Осуществляет поиск необходимой информации, хранящейся в структуре больших данных, опираясь на результаты анализа поставленной задачи;

Разрабатывает варианты решения проблемной ситуации на основе критического анализа доступных источников информации в структуре больших данных;

Использует основные методы, средства получения, представления, хранения и обработки статистических данных с использованием методов и технологий больших данных;

Применяет статистические методы обработки собранных данных, использует анализ данных, необходимых для решения поставленных экономических задач на основе больших данных;

Использует соответствующие содержанию профессиональных задач современные цифровые информационные технологии, основываясь на принципах их работы, в том числе, на принципах обработки больших данных;

Понимает принципы работы современных цифровых информационных технологий, соответствующих содержанию профессиональных задач на основе методов и принципов хранения, выборки и обработки больших данных.

2. Место дисциплины в структуре ОПОП ВО

Для освоения дисциплины необходимы навыки программирования и знания основ анализа данных

3. Планируемые результаты обучения по дисциплине

Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения ОПОП (компетенциями и индикаторами достижения компетенций)

Формируемые компетенции (код, содержание компетенции)	Планируемые результаты обучения по дисциплине, в соответствии с индикатором достижения компетенции		Наименование оценочного средства
	Индикатор достижения компетенции	Результаты обучения по дисциплине	
ПК-2 Владение методами программной реализации распределенных информационных систем	ПК-2.1 Реализует методы и программные интерфейсы взаимодействия с внешними программными компонентами	методы и программные интерфейсы взаимодействия с внешними программными компонентами (ПК-2.1) принципы работы соответствующих содержанию профессиональных задач современных цифровых информационных технологий (ПК-2.1) Реализовывать методы и программные интерфейсы взаимодействия с внешними программными	вопросы к устному опросу

		компонентами (ПК-2.1) Методами и программными интерфейсами взаимодействия с внешними программными компонентами (ПК-2.1) навыками создания программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов (ПК-2.1)	
ПК-3 Владение навыками создания программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов	ПК-3.2 Применяет методы машинного обучения для обработки информации	методы машинного обучения для обработки информации (ПК-3.2) Применять методы машинного обучения для обработки информации (ПК-3.2) методами машинного обучения для обработки информации больших данных с использованием алгоритмов обработки и анализа (ПК-3.2)	

4. Структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 7 зачетных единиц, 252 час.

4.1. Форма обучения: очная

Уровень базового образования: высшее.
Срок обучения 2г.

4.1.1. Структура дисциплины

№ п\п	Раздел (тема) дисциплины	Семестр	Контактная работа обучающихся с педагогическим работником							Самостоятельная работа	Форма текущего контроля успеваемости (по неделям семестра), форма промежуточной аттестации(по семестрам)
			Лекции	Практические занятия	Лабораторные работы	Контрольные работы	КП / КР	Консультация	Контроль		
1	Основы работы с большими данными	1	12							45	Устный опрос
2	Apache Spark с Python	1			24					41,55	Устный опрос
Всего за семестр		126	12		24			1,2	2,25	86,55	Зач.
3	Apache Spark с Python	2	20	14	36					23	Устный опрос
Всего за семестр		126	20	14	36		+	4	2,35	23	Экз.(26,65)
Итого		252	32	14	60			5,2	4,6	109,55	26,65

4.1.2. Содержание дисциплины

4.1.2.1. Перечень лекций

Семестр 1

Раздел 1. Основы работы с большими данными

Лекция 1.

Определение больших данных (2 часа).

Лекция 2.

Метаданные (2 часа).

Лекция 3.

Большие данные (2 часа).

Лекция 4.

Системы управления.Большими данными (2 часа).

Лекция 5.

Программная платформа Hadoop.Распределенная файловая система HDFS (2 часа).

Лекция 6.

Apache Spark с Python (2 часа).

Семестр 2
Раздел 3. Apache Spark с Python
Лекция 7.
Работа со SparkContext (2 часа).
Лекция 8.
PySpark SQL (2 часа).
Лекция 9.
Apache Spark. RDD, DataFrame, DataSet (2 часа).
Лекция 10.
Apache Spark Streaming (2 часа).
Лекция 11.
Apache Spark Structured Streaming (2 часа).
Лекция 12.
Apache Spark (2 часа).
Лекция 13.
User-Defined Function (UDF) (2 часа).
Лекция 14.
Сбор данных на потоке Apache Kafka (2 часа).
Лекция 15.
Перенос моделей в распределенную среду (2 часа).
Лекция 16.
Тестирование моделей (2 часа).

4.1.2.2. Перечень практических занятий

Семестр 2
Раздел 3. Apache Spark с Python
Практическое занятие 1
Получение и загрузка датафрейма (2 часа).
Практическое занятие 2
Работа с датафреймом (2 часа).
Практическое занятие 3
Разметка данных (2 часа).
Практическое занятие 4
Разведочный анализ данных (2 часа).
Практическое занятие 5
Apache Spark. RDD, DataFrame, DataSet (2 часа).
Практическое занятие 6
Apache Spark Streaming (2 часа).
Практическое занятие 7
Apache Spark Structured Streaming (2 часа).

4.1.2.3. Перечень лабораторных работ

Семестр 1
Раздел 2. Apache Spark с Python
Лабораторная 1.
PySpark Machine learning (4 часа).
Лабораторная 2.
PySpark RDD continued (4 часа).
Лабораторная 3.
MPI для распределенного решения задачи кластеризации и оценки качества кластеризации (4 часа).
Лабораторная 4.
Введение в облачные технологии (4 часа).

Лабораторная 5.
Вычисления на графическом ядре (4 часа).
Лабораторная 6.
PySpark в облачном кластере (4 часа).
Семестр 2
Раздел 3. Apache Spark с Python
Лабораторная 7.
PySpark и SQL (4 часа).
Лабораторная 8.
PySpark и SQL (4 часа).
Лабораторная 9.
Интеграция Apache Spark с Python (4 часа).
Лабораторная 10.
Использование SparkContext (4 часа).
Лабораторная 11.
Работа с компонентами Apache Spark: RDD (4 часа).
Лабораторная 12.
Работа с компонентами Apache Spark: DataFrame (4 часа).
Лабораторная 13.
Работа с компонентами Apache Spark: DataSet (4 часа).
Лабораторная 14.
Apache Spark Streaming (4 часа).
Лабораторная 15.
Apache Spark Structured Streaming (4 часа).

4.1.2.4. Перечень тем и учебно-методическое обеспечение самостоятельной работы

Перечень тем, вынесенных на самостоятельное изучение:
1. Архитектура Apache Spark: Основные компоненты Spark.
2. Операции над RDD (трансформации и действия).
3. Использование SQL-запросов для анализа данных.
4. Работа с Hive и интеграция с другими источниками данных.
5. Обработка потоковых данных с Spark Streaming.
6. Алгоритмы и применение MLlib для построения моделей.
7. Графовые вычисления с GraphX.
8. Работа с графами и алгоритмами на графах.
9. Оптимизация производительности Spark.
10. Интеграция Spark с Hadoop.
11. Чтение и запись данных из различных форматов (CSV, JSON, Parquet и др.), интеграция с базами данных.
12. Сравнение Apache Spark с Hadoop MapReduce, Flink и другими инструментами.

Для самостоятельной работы используются методические указания по освоению дисциплины и издания из списка приведенной ниже основной и дополнительной литературы.

4.1.2.5. Перечень тем контрольных работ, рефератов, ТР, РГР, РПР

Не планируется.

4.1.2.6. Примерный перечень тем курсовых работ (проектов)

1. Разработка распределенного механизма запросов с низкой задержкой для больших наборов данных, включая структурированные и полуструктурированные/вложенные данные.
2. Разработать инфраструктуру хранилища данных на базе SQL для чтения, записи и управления большими датасетами в распределенных средах хранения.
3. Разработка системы дата-пайплайна и упрощение инкрементной обработки данных.

4. Разработать механизм аналитической обработки информации, предназначенный для работы с очень большими массивами данных.
5. Разработать аналитическую базу данных реального времени с собственной поддержкой вложенных данных с использованием нативных инвертированных поисковых индексов.

5. Образовательные технологии

В процессе изучения дисциплины применяется контактная технология преподавания (за исключением самостоятельно изучаемых студентами вопросов). При проведении практических работ применяется имитационный или симуляционный подход. Шаги решения задач студентам демонстрируются при помощи мультимедийной техники. В дальнейшем студенты самостоятельно решают аналогичные задания.

6. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины.

Фонды оценочных материалов (средств) приведены в приложении.

7. Учебно-методическое и информационное обеспечение дисциплины.

7.1. Основная учебно-методическая литература по дисциплине

1. Целых, А. Н. Принятие решений на основе методов машинного обучения : учебное пособие по курсам «Модели и методы инженерии знаний», «Методы анализа больших данных» / А. Н. Целых, Н. В. Драгныш, Э. М. Котов. — Ростов-на-Дону, Таганрог : Издательство Южного федерального университета, 2022. — 113 с. — ISBN 978-5-9275-4246-8. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/131458.html. — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/131458.html

2. Конкина, В. В. Введение в большие данные и анализ информации : учебное пособие / В. В. Конкина, А. Б. Борисенко, И. Л. Коробова. — Тамбов : Тамбовский государственный технический университет, ЭБС АСВ, 2024. — 81 с. — ISBN 978-5-8265-2749-8. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/145326.html. — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/145326.html

3. Параскевов, А. В. Большие данные : учебник / А. В. Параскевов, А. Э. Сергеев. — Москва, Вологда : Инфра-Инженерия, 2024. — 148 с. — ISBN 978-5-9729-2120-1. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/143597.html. — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/143597.html

7.2. Дополнительная учебно-методическая литература по дисциплине

1. Железнов, М. М. Методы и технологии обработки больших данных : учебно-методическое пособие / М. М. Железнов. — Москва : МИСИ-МГСУ, ЭБС АСВ, 2020. — 46 с. — ISBN 978-5-7264-2193-3. — Текст : электронный // Цифровой образовательный ресурс IPR SMART : [сайт]. — URL: https://www.iprbookshop.ru/101802.html. — Режим доступа: для авторизир. пользователей - https://www.iprbookshop.ru/101802.html

7.3. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень программного обеспечения и информационных справочных систем

В образовательном процессе используются информационные технологии, реализованные на основе информационно-образовательного портала института (www.mivlgu.ru/iop), и инфокоммуникационной сети института:

- предоставление учебно-методических материалов в электронном виде;

- взаимодействие участников образовательного процесса через локальную сеть института и Интернет;

- предоставление сведений о результатах учебной деятельности в электронном личном кабинете обучающегося.

Информационные справочные системы:

https://sparkbyexamples.com/pyspark/

Программное обеспечение:

LibreOffice (Mozilla Public License v2.0)

7-Zip (GNU LGPL)

Google Chrome (Лицензионное соглашение Google)

РЕД ОС (Соглашение №140/05-21У от 18.05.2021 года о сотрудничестве в области науки, развития инновационной деятельности)

Microsoft Visual Studio (Программа Microsoft Azure Dev Tools for Teaching (Order Number: IM126433))

Microsoft Visio (Программа Microsoft Azure Dev Tools for Teaching (Order Number: IM126433))

Notepad++ (GNU GPL 3)

7.4. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

iprbookshop.ru

mivlgu.ru/iop

8. Материально-техническое обеспечение дисциплины

Лаборатория системного и прикладного программирования

6 шт. компьютеров Intel Core i5, 3500 MHz/ ОЗУ 6Gb/ SSD-512Gb/ LG 22’; 6 шт. персональных компьютеров Digitech (комплект2) Intel Core i5 3000 MHz/ DDR-4 12Gb/ SSD-512Gb/ Philips 21eb; проектор NEC V300X 3D; экран проекционный настенный Lumien Master Picture; маршрутизатор Gigabit Switch TEG-S16S; макет системы мобильного мониторинга; лабораторный стенд для изучения микроконтроллера; роботизированная платформа IE-POP-BOT; аппаратно-программный комплекс «Изучение принципов построения и исследования инфокоммутационных локальных сетей». Маркерная доска. Доступ к сети Интернет.

Лаборатория информационных ресурсов

6 шт. компьютеров Intel Core i5, 3500 MHz/ ОЗУ 6 Gb/ SSD-512Gb/ LG 22’; 6 шт. персональных компьютеров Digitech (комплект 2) Intel Core i5 3000 MHz/ DDR-4 12Gb/ SSD-512Gb/ Philips 21eb; проектор NEC V300X 3D; экран проекционный настенный Lumien Master Picture; маршрутизатор Gigabit Switch TEG-S16S; макет системы мобильного мониторинга; лабораторный стенд для изучения микроконтроллера; роботизированная платформа IE-POP-BOT; аппаратно-программный комплекс «Изучение принципов построения и исследования инфокоммутационных локальных сетей». Маркерная доска. Доступ к сети Интернет.

9. Методические указания по освоению дисциплины

Для успешного освоения теоретического материала обучающийся: знакомится со списком рекомендуемой основной и дополнительной литературы; уточняет у преподавателя, каким дополнительным пособиям следует отдать предпочтение; ведет конспект лекций и прорабатывает лекционный материал, пользуясь как конспектом, так и учебными пособиями.

На практических занятиях пройденный теоретический материал подкрепляется решением задач по основным темам дисциплины. Занятия проводятся в компьютерном классе, используя специальное программное обеспечение. Каждой подгруппе обучающихся преподаватель выдает задачу, связанную с разработкой и программной реализацией алгоритмов обработки информации. В конце занятия обучающие демонстрируют полученные результаты преподавателю и при необходимости делают работу над ошибками.

До выполнения лабораторных работ обучающийся изучает соответствующий раздел теории. Перед занятием студент знакомится с описанием заданий для выполнения работы, внимательно изучает содержание и порядок проведения лабораторной работы. Лабораторная работа проводятся в компьютерном классе. Обучающиеся выполняют индивидуальную задачу компьютерного моделирования в соответствии с заданием на лабораторную работу. Полученные результаты исследований сводятся в отчет и защищаются по традиционной методике в классе на следующем лабораторном занятии. Необходимый теоретический материал, индивидуальное задание, шаги выполнения лабораторной работы и требование к отчету приведены в методических указаниях, размещенных на информационно-образовательном портале института.

Самостоятельная работа оказывает важное влияние на формирование личности будущего специалиста, она планируется обучающимся самостоятельно. Каждый обучающийся самостоятельно определяет режим своей работы и меру труда, затрачиваемого на овладение учебным содержанием дисциплины. Он выполняет внеаудиторную работу и изучение разделов, выносимых на самостоятельную работу, по личному индивидуальному плану, в зависимости от его подготовки, времени и других условий.

Курсовая работа выполняется в соответствии с методическими указаниями на курсовую работу. Обучающийся выбирает одну из указанных в перечне тем курсовых работ, исходя из своих интересов, наличия соответствующих литературных и иных источников. В ходе выполнения курсовой работы преподаватель проводит консультации обучающегося. На заключительном этапе обучающийся оформляет пояснительную записку к курсовой работе и выполняет ее защиту в присутствии комиссии из преподавателей кафедры.

Форма заключительного контроля при промежуточной аттестации – экзамен. Для проведения промежуточной аттестации по дисциплине разработаны фонд оценочных средств и балльно-рейтинговая система оценки учебной деятельности студентов. Оценка по дисциплине выставляется в информационной системе и носит интегрированный характер, учитывающий результаты оценивания участия студентов в аудиторных занятиях, качества и своевременности выполнения заданий в ходе изучения дисциплины и промежуточной аттестации.

Программа составлена в соответствии с требованиями ФГОС ВО по направлению *09.04.04 Программная инженерия* и профилю подготовки *Технологии разработки интеллектуальных систем*
Рабочую программу составил *Мортин К.В.*_____

Программа рассмотрена и одобрена на заседании кафедры *ПИн*

протокол № 10 от 14.05.2024 года.
Заведующий кафедрой *ПИн* _____*Жизняков А.Л.*
(Подпись)

Рабочая программа рассмотрена и одобрена на заседании учебно-методической комиссии факультета

протокол № 9 от 17.05.2024 года.
Председатель комиссии ФИТР _____ *Рыжкова М.Н.*
(Подпись) (Ф.И.О.)

Приложение

Фонд оценочных материалов (средств) по дисциплине
Системы обработки больших данных

1. Оценочные материалы для проведения текущего контроля успеваемости по дисциплине

Темы для устного опроса:

1. Что такое большие данные (Big Data)?

Опишите основные характеристики больших данных, такие как объем, скорость и разнообразие.

2. Что такое Apache Spark и как он отличается от Hadoop?

Объясните архитектуру Spark и его преимущества по сравнению с традиционными решениями на основе Hadoop.

3. Что такое RDD (Resilient Distributed Dataset) в Spark?

Опишите, что такое RDD, его основные свойства и как он используется для обработки данных.

4. Как работает механизм ленивых вычислений в Spark?

Объясните, что такое ленивые вычисления и как они помогают оптимизировать выполнение задач в Spark.

5. Что такое DataFrame и как он отличается от RDD?

Опишите структуру DataFrame, его преимущества и случаи использования по сравнению с RDD.

6. Как можно использовать Spark SQL для анализа данных?

Приведите примеры, как использовать Spark SQL для выполнения запросов к данным, находящимся в различных источниках.

7. Что такое трансформации и действия (actions) в Spark?

Объясните разницу между трансформациями и действиями, приведите примеры каждого типа.

8. Как Spark обрабатывает данные в режиме реального времени?

Опишите, как работает Spark Streaming и какие сценарии применения он поддерживает.

9. Что такое кластеризация в контексте Spark?

Объясните, как Spark распределяет задачи по узлам кластера и какие компоненты участвуют в этом процессе.

10. Каковы основные способы оптимизации производительности приложений на Spark?

Перечислите и опишите несколько методов оптимизации производительности, таких как использование кэширования, выбор подходящего уровня параллелизма и оптимизация запросов.

Общее распределение баллов текущего контроля по видам учебных работ для студентов

Рейтинг-контроль 1	Устный опрос	20
Рейтинг-контроль 2	Устный опрос	20
Рейтинг-контроль 3	Устный опрос	20
Посещение занятий студентом		0
Дополнительные баллы (бонусы)		0
Выполнение семестрового плана самостоятельной работы		0

2. Промежуточная аттестация по дисциплине

Перечень вопросов к экзамену / зачету / зачету с оценкой.
Перечень практических задач / заданий к экзамену / зачету / зачету с оценкой (при наличии)

Методические материалы, характеризующие процедуры оценивания

https://www.mivlgu.ru/iop/course/view.php?id=3734

Максимальная сумма баллов, набираемая студентом по дисциплине равна 100.

Оценка в баллах	Оценка по шкале	Обоснование	***Уровень сформированности компетенций***
Более 80	«Отлично»	Содержание курса освоено полностью, без пробелов, необходимые практические навыки работы с освоенным материалом сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено числом баллов, близким к максимальному	***Высокий уровень***
66-80	«Хорошо»	Содержание курса освоено полностью, без пробелов, некоторые практические навыки работы с освоенным материалом сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками	***Продвинутый уровень***

50-65	«Удовлетворительно»	Содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий, возможно, содержат ошибки	***Пороговый уровень***
Менее 50	«Неудовлетворительно»	Содержание курса не освоено, необходимые практические навыки работы не сформированы, выполненные учебные задания содержат грубые ошибки	***Компетенции не сформированы***

3. Задания в тестовой форме по дисциплине

Примеры заданий:

Что такое "большие данные" (Big Data)?

A) Данные, которые не помещаются на одном жестком диске

B) Данные, которые обрабатываются с помощью обычных реляционных баз данных

C) Данные, которые могут быть обработаны вручную

D) Данные, размер которых превышает 1 ТБ

Вопрос 2:

Какой из следующих компонентов является частью экосистемы Apache Spark?

A) Hadoop HDFS

B) MySQL

C) MongoDB

D) Microsoft Excel

Вопрос 3:

Какой из следующих методов используется для обработки данных в Spark?

A) MapReduce

B) DataFrame

C) SQL Server

D) OLAP

Вопрос 4:

Какой из следующих языков программирования поддерживает написание приложений на Apache Spark?

A) Java

B) C#

C) Ruby

D) Все перечисленные

Вопрос 5:

Что такое RDD в Apache Spark?

A) Реляционная база данных

B) Диск с данными

C) Устойчивый распределенный набор данных

D) Виртуальная машина

Полный перечень тестовых заданий с указанием правильных ответов, размещен в банке вопросов на информационно-образовательном портале института по ссылке https://www.mivlgu.ru/iop/course/view.php?id=3734

Оценка рассчитывается как процент правильно выполненных тестовых заданий из их общего числа.