Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

Кафедра математики и цифровых технологий

РАБОЧАЯ ПРОГРАММА

ДИСЦИПЛИНЫ

«Б1.Д.В.9 Компьютерные технологии обработки больших массивов данных»

Уровень высшего образования

БАКАЛАВРИАТ

Направление подготовки

01.03.02 Прикладная математика и информатика
(код и наименование направления подготовки)

Прикладное программирование и корпоративные информационные системы
(наименование направленности (профиля) образовательной программы)

Квалификация

Бакалавр

Форма обучения

Очная

Год набора 2024

Рабочая программа дисциплины «*Б1.Д.В.9 Компьютерные технологии обработки больших массивов данных*» рассмотрена и утверждена на заседании кафедры

Кафедра математики и цифровых технологий
наименование кафедры

протокол № 6 от " 19 " февраля 2024 г.

Заведующий кафедрой
Кафедра математики и цифровых технологий _____ А.Е. Шухман
наименование кафедры *подпись* *расшифровка подписи*

Исполнитель:
Доцент кафедры МЦТ _____ Ю.А.Ушаков
должность *подпись* *расшифровка подписи*

СОГЛАСОВАНО:
Председатель методической комиссии по направлению подготовки
01.03.02 Прикладная математика и информатика _____ И.П. Болодурина
код наименование *личная подпись* *расшифровка подписи*

Заведующий отделом формирования фонда и научной обработки документов
_____ Н.Н. Бигалиева
личная подпись *расшифровка подписи*

Уполномоченный по качеству института
_____ И.В. Крючкова
личная подпись *расшифровка подписи*

№ регистрации _____

© Ушаков Ю.А., 2024
© ОГУ, 2024

1 Цели и задачи освоения дисциплины

Цель (цели) освоения дисциплины:

Ознакомление с основными технологиями решения задач обработки больших по объему, быстро изменяющихся и плохо структурированных данных;

Задачи:

- формирование представлений о современных методах решения задач обработки больших данных;
- формирование умения применять методы анализа больших данных;
- формирование умения реализовывать приложения для аналитики больших данных.

2 Место дисциплины в структуре образовательной программы

Дисциплина относится к обязательным дисциплинам (модулям) вариативной части блока Д «Дисциплины (модули)»

Пререквизиты дисциплины: *Б1.Д.Б.18 Языки программирования, Б1.Д.Б.24 Теория вероятностей и математическая статистика, Б1.Д.Б.26 Базы данных, Б1.Д.Б.31 Моделирование информационных систем, Б1.Д.В.2 Технология программирования*

Постреквизиты дисциплины: *Б2.П.В.П.2 Преддипломная практика, ФДТ.2 Методология кибериммунитета*

3 Требования к результатам обучения по дисциплине

Процесс изучения дисциплины направлен на формирование следующих результатов обучения

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
ПК*-1 Способен собирать, обрабатывать и интерпретировать результаты современных научных исследований, а также проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК*-1-В-1 Имеет представление о приемах и способах сбора, методах и алгоритмах обработки и интерпретации материалов современных научных исследований, необходимых для формирования выводов по соответствующим научным проблемам ПК*-1-В-2 Собирает и обрабатывает статистический, экспериментальный, теоретический, графический и т.п. материал, необходимый для построения математических моделей, расчетов и конкретных практических выводов ПК*-1-В-3 Выбирает методы проведения исследований и осуществляет качественный и количественный анализ собранных данных в конкретной области профессиональной деятельности	**Знать:** принципы программно-алгоритмической обработки больших данных; **Уметь:** разрабатывать алгоритмы и программы обработки больших данных; формировать программное обеспечение **Владеть:** навыками программной реализации алгоритмов обработки больших данных;
ПК*-3 Способен применять современные технологии анализа данных в информационно-	ПК*-3-В-1 Имеет представление о способах получения, обработки, передачи и хранения данных в информационно-аналитической деятельности при	**Знать:** способы хранения, обработки и представления данных в

Код и наименование формируемых компетенций	Код и наименование индикатора достижения компетенции	Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций
аналитической деятельности систем управления и принятия решений, а также для разработки на их основе новых продуктов и услуг	использовании различных систем управления ПК*-3-В-2 Использует современные методы управления стратегическим развитием методологической и технологической инфраструктуры анализа данных в информационно-аналитической деятельности ПК*-3-В-3 Разрабатывает инфраструктурные решения, сервисы, продукты и услуги на основе анализа данных информационных систем ПК*-3-В-4 Демонстрирует навыки принятия решений на основе анализа данных современных информационных систем управления	распределенных **Уметь:** выбирать и применять методы конфигурирования информационных систем и оптимизации баз данных **Владеть:** навыками использования актуальных инструментальных средств для разработки программного обеспечения

4 Структура и содержание дисциплины

4.1 Структура дисциплины

Общая трудоемкость дисциплины составляет 8 зачетных единиц (288 академических часов).

Вид работы	Трудоемкость, академических часов		
	7 семестр	8 семестр	всего
Общая трудоёмкость	**180**	**108**	**288**
Контактная работа:	**51,25**	**41,25**	**92,5**
Лекции (Л)	34	24	58
Лабораторные работы (ЛР)	16	16	32
Консультации	1	1	2
Промежуточная аттестация (зачет, экзамен)	0,25	0,25	0,5
Самостоятельная работа: *- самоподготовка: проработка и повторение лекционного материала и материала учебников и учебных пособий;* *- изучение разделов курса в системе электронного обучения;* *- подготовка к лабораторным занятиям;* *- подготовка к рубежному контролю*	**128,75**	**66,75**	**195,5**
Вид итогового контроля (зачет, экзамен, дифференцированный зачет)	**экзамен**	**экзамен**	

Разделы дисциплины, изучаемые в 7 семестре

№ раздела	Наименование разделов	Количество часов				
		всего	аудиторная работа			внеауд. работа
			Л	ПЗ	ЛР	

№ раздела	Наименование разделов	Количество часов				
		всего	аудиторная работа			внеауд. работа
			Л	ПЗ	ЛР	
1	Введение в «большие данные»	54	10		4	40
2	Обработка потоков больших данных	60	12		8	40
3	Алгоритмы кластеризации больших многомерных наборов данных	66	12		4	50
	Итого:	180	34		16	130

Разделы дисциплины, изучаемые в 8 семестре

№ раздела	Наименование разделов	Количество часов				
		всего	аудиторная работа			внеауд. работа
			Л	ПЗ	ЛР	
4	Фреймворк Apache Spark как способ обработки больших объёмов данных	26	6		4	16
5	Анализ массивов и временных рядов	26	6		4	16
6	Вопросы хранения больших данных	26	6		4	16
7	Извлечение, сбор и валидация Big Data	30	6		4	20
	Итого:	108	24		16	68
	Всего:	288	58		32	198

4.2 Содержание разделов дисциплины

1 Введение в «большие данные». История возникновения термина «Большие данные». Источники больших данных. Распределенные файловые системы и технология распределения-редукции. Группировка по ключу. Детали выполнения MapReduce. Комбинаторы. Обработка отказов узлов..

2 Обработка потоков больших данных. Специализированные алгоритмы для работы с данными. Система управления потоками данных. Запросы к потокам. Проблемы обработки потоков. Выборка данных из потоков. Фильтрация потоков. Алгоритм Флажоле-Мартена. Оценивание моментов. Алгоритм Алона-Матиаса-Сегеди для вторых моментов.

3 Алгоритмы кластеризации больших многомерных наборов данных. Иерархическая кластеризация. Альтернативные правила управления иерархической кластеризацией. Алгоритм к-средних. Обработка данных в алгоритме BFR. Алгоритм CURE. Кластеризация для потоков и параллелизм.

4 Фреймворк Apache Spark как способ обработки больших объёмов данных. **Г**оризонтальная масштабируемость. Основная технология распараллеливания. Кластеры среды hadoop. Особенности системы хранения. Главные конкуренты и вычислительные мощность

5 Анализ массивов и временных рядов. Rolling window estimations. Экспоненциальное сглаживание, модель ХольтаВинтерса. Кросс-валидация на временных рядах, подбор параметров. Линейные модели на временных рядах. Извлечение признаков (Feature extraction). Линейная регрессия vs XGBoost.

6 Вопросы хранения больших данных. Вопросы безопасности больших данных. Резервное копирование облака. Собственные алгоритмы шифрования. Типичные проблемы

7 Извлечение, сбор и валидация Big Data. Определение проблемы. Извлечение данных. Подготовка данных — очистка данных, преобразование данных. Предсказательные модели.

4.3 Лабораторные работы

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов

№ ЛР	№ раздела	Наименование лабораторных работ	Кол-во часов
1	1	Источники больших данных. Знакомство с примерами прикладных систем добычи данных. Принцип Бонферрони.	2
2	1	Задачи распределители. Группировка по ключу. Детали выполнения MapReduce. Комбинаторы. Обработка отказов узлов	2
3	2	Операции реляционной алгебры. Вычисление выборки с помощью MapReduce. Разбиение документов на шинглы. Хэширование документов с учетом близости. Методы для высокой степени сходства.	4
4	2	Обработка потоков данных и специализированные алгоритмы для работы с данными.	4
5	3	Система управления потоками данных. Запросы к потокам. Проблемы обработки потоков. Выборка данных из потоков. Фильтрация потоков. Алгоритм Флажоле-Мартена. Оценивание моментов. Алгоритм Алона-Матиаса-Сегеди для вторых моментов.	4
6	4	Получение репрезентативной выборки. Динамическое изменение размера выборки. Анализ фильтра Блума. Комбинирование оценок. Обработка бесконечных потоков.	4
7	5	Алгоритмы кластеризации больших многомерных наборов данных. Иерархическая кластеризация. Альтернативные правила управления иерархической кластеризацией. Алгоритм к-средних	4
8	6	Алгоритмы машинного обучения, применяемые к большим наборам данных. Перцептроны. Механизм метода опорных векторов. Нахождение оптимальных приближений разделителей. Обучение по ближайшим соседям	4
9	7	Обучающие наборы. Обучение перцептрона с нулевым порогом. Многоклассовые перцептроны. Параллельная реализация перцептронов. Нормировка гиперплоскости. Нахождение решений в методе опорных векторов с помощью градиентного спуска.	4
		Итого:	32

5 Учебно-методическое обеспечение дисциплины

1. Гергель, В. П. Современные языки и технологии параллельного программирования [Текст] : учебное пособие для студентов высших учебных заведений, обучающихся по направлениям ВПО 010400 "Прикладная математика и информатика" и 010300 "Фундаментальная информатика и информационные технологии" / В. П. Гергель; Б-ка Нижегор. гос. ун-та им. Н. И. Лобачевского. - Москва : Изд-во Моск. ун-та, 2012. - 407 с.

2. Антонов, А. С. Технологии параллельного программирования MPI и OpenMP [Текст] : учебное пособие для студентов высших учебных заведений, обучающихся по направлениям ВПО 010400 "Прикладная математика и информатика" и 010300 "Фундаментальная информатика и информацион- ные технологии" / А. С. Антонов; Моск. гос. ун-т им. М. В. Ломоносова. - Москва : Изд-во Моск. ун- та, 2012. - 340 с.

5.2 Дополнительная литература

1. Воеводин В.В., Воеводин Вл.В. Параллельные вычисления. – СПб: БХВ-Петербуг, 2002. – 608 с.

2. Гергель, В. П. Теория и практика параллельных вычислений : учебное пособие / В. П. Гергель. – Москва : Интернет-Университет Информационных Технологий (ИНТУИТ) : Бином. Лаборатория знаний, 2007. – 424 с. : ил.,табл. – (Основы информационных технологий). – Режим

доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=233067 . – ISBN 978-5-9556-0096-3. – Текст : электронный.

3. Инструменты параллельного программирования в системах с общей памятью [Текст] : учеб. для вузов / К. В. Корняков [и др.]; под ред. В. П. Гергеля ; Нижегород. гос. ун-т им. Н. И. Лоба- чевского.- 2-е изд., испр. и доп. - М. : Изд-во Моск. ун-та, 2010. - 267 с.

5.3 Интернет-ресурсы

https://openedu.ru/course/hse/big_data_processing/ - «Открытое образование», Каталог курсов, МООК: «Обработка больших данных»;

https://openedu.ru/course/misis/ABD/?session=spring_2023 - «Открытое образование», Каталог курсов, МООК: «Введение в инженерию больших данных».

5.4 Программное обеспечение, профессиональные базы данных и информационные справочные системы

1. Операционная система РЕД ОС для рабочих станций, имеется лицензия, входит в реестр отечественного ПО.
2. «МойОфис Образование» – набор приложений для работы с текстом, таблицами и презентациями в образовательных организациях
3. Система управления учебным процессом Moodle, свободно распространяемая.
4. Программная система для организации видео-конференц-связи MTS Link
5. Программа для просмотра сайтов Яндекс.Браузер, свободно распространяемая, входит в реестр отечественного ПО.

6 Материально-техническое обеспечение дисциплины

Учебные аудитории для проведения занятий лекционного типа, семинарского типа, для проведения групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации оснащены комплектами ученической мебели, техническими средствами обучения, служащими для представления учебной информации большой аудитории. Для чтения лекций используется переносной мультимедийный комплект: ноутбук, проектор, экран.

Для проведения лабораторных занятий используются компьютерные классы, оснащенные компьютерами, объединенными в локальную сеть, подключенную через университетскую сеть к сети Интернет. Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой, подключенной к сети "Интернет", и обеспечением доступа в электронную информационно-образовательную среду ОГУ.

К рабочей программе прилагаются:

– Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине;

– Методические указания для обучающихся по освоению дисциплины.