

Минобрнауки России

Федеральное государственное бюджетное образовательное учреждение  
высшего образования

**«Оренбургский государственный университет»**

Кафедра геометрии и компьютерных наук

## **РАБОЧАЯ ПРОГРАММА**

### **ДИСЦИПЛИНЫ**

*«Б1.Д.В.Э.3.1 Технологии обработки больших данных»*

Уровень высшего образования

**БАКАЛАВРИАТ**

Направление подготовки

**09.03.02 Информационные системы и технологии**  
(код и наименование направления подготовки)

**Анализ данных и машинное обучение**

(наименование направленности (профиля) образовательной программы)

Квалификация

**Бакалавр**

Форма обучения

**Очная**

Год набора 2023

Рабочая программа дисциплины «Б1.Д.В.Э.3.1 Технологии обработки больших данных» рассмотрена и утверждена на заседании кафедры

Кафедра геометрии и компьютерных наук  
*наименование кафедры*

протокол № 6 от "17" февраля 2023 г.

Заведующий кафедрой

Кафедра геометрии и компьютерных наук  
*наименование кафедры*

  
*подпись*

А.Е. Шухман  
*расшифровка подписи*

Исполнитель:

Доцент кафедры ГКН  
*должность*

  
*подпись*

Ю.А. Ушаков  
*расшифровка подписи*

СОГЛАСОВАНО:

Председатель методической комиссии по направлению подготовки

09.03.02 Информационные системы и технологии  
*код наименование*

  
*личная подпись*

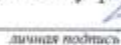
А.Е. Шухман  
*расшифровка подписи*

Заведующий отделом комплектования научной библиотеки

  
*личная подпись*

Н.Н. Бигалиева  
*расшифровка подписи*

Уполномоченный по качеству факультета

  
*личная подпись*

И.В. Крючкова  
*расшифровка подписи*

№ регистрации \_\_\_\_\_

## 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 дисциплинам (модулям) по выбору вариативной части блока Д «Дисциплины (модули)»

Пререквизиты дисциплины: *Б1.Д.Б.18 Теория вероятностей и математическая статистика, Б1.Д.Б.25 Инструментальные средства информационных систем, Б1.Д.Б.26 Технологии обработки информации, Б1.Д.В.7 Моделирование процессов и систем*

Постреквизиты дисциплины: *Отсутствуют*

## 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*-3 Способен проектировать системы сбора, анализа, обработки и представления разнородных данных                                              | ПК*-3-В-3 Применяет технологии распределенной обработки больших данных                                                       | <b><u>Знать:</u></b><br>принципы программно-алгоритмической обработки больших данных;<br><b><u>Уметь:</u></b><br>разрабатывать алгоритмы и программы обработки больших данных;<br>формировать программное обеспечение<br><b><u>Владеть:</u></b><br>навыками программной реализации алгоритмов обработки больших данных; |
| ПК*-6 Способен обеспечивать развертывание, сопровождение и оптимизацию функционирования баз данных, являющихся частью различных информационных | ПК*-6-В-2 Конфигурирует информационные системы и оптимизирует базы данных, являющиеся частью различных информационных систем | <b><u>Знать:</u></b><br>способы хранения, обработки и представления данных в распределенных<br><b><u>Уметь:</u></b><br>выбирать и применять                                                                                                                                                                             |

| Код и наименование формируемых компетенций | Код и наименование индикатора достижения компетенции | Планируемые результаты обучения по дисциплине, характеризующие этапы формирования компетенций                                                                                                |
|--------------------------------------------|------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| систем                                     |                                                      | методы конфигурирования информационных систем и оптимизации баз данных<br><b>Владеть:</b> навыками использования актуальных инструментальных средств для разработки программного обеспечения |

## 4 Структура и содержание дисциплины

### 4.1 Структура дисциплины

Общая трудоемкость дисциплины составляет 4 зачетные единицы (144 академических часа).

| Вид работы                                                                                                                                                                                                                                    | Трудоемкость, академических часов |               |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|---------------|
|                                                                                                                                                                                                                                               | 8 семестр                         | всего         |
| <b>Общая трудоёмкость</b>                                                                                                                                                                                                                     | <b>144</b>                        | <b>144</b>    |
| <b>Контактная работа:</b>                                                                                                                                                                                                                     | <b>34,25</b>                      | <b>34,25</b>  |
| Лекции (Л)                                                                                                                                                                                                                                    | 14                                | 14            |
| Лабораторные работы (ЛР)                                                                                                                                                                                                                      | 20                                | 20            |
| Промежуточная аттестация (зачет, экзамен)                                                                                                                                                                                                     | 0,25                              | 0,25          |
| <b>Самостоятельная работа:</b><br><i>самоподготовка (проработка и повторение лекционного материала и материала учебников и учебных пособий);</i><br><i>- подготовка к лабораторным занятиям;</i><br><i>- подготовка к рубежному контролю.</i> | <b>109,75</b>                     | <b>109,75</b> |
| <b>Вид итогового контроля (зачет, экзамен, дифференцированный зачет)</b>                                                                                                                                                                      | <b>зачет</b>                      |               |

Разделы дисциплины, изучаемые в 8 семестре

| № раздела | Наименование разделов                                      | Количество часов |                   |    |    |                |
|-----------|------------------------------------------------------------|------------------|-------------------|----|----|----------------|
|           |                                                            | всего            | аудиторная работа |    |    | внеауд. работа |
|           |                                                            |                  | Л                 | ПЗ | ЛР |                |
| 1         | Введение в «большие данные»                                | 44               | 2                 |    | 6  | 36             |
| 2         | Обработка потоков больших данных                           | 46               | 4                 |    | 6  | 36             |
| 3         | Алгоритмы кластеризации больших многомерных наборов данных | 54               | 8                 |    | 8  | 38             |
|           | Итого:                                                     | 144              | 14                |    | 20 | 110            |
|           | Всего:                                                     | 144              | 14                |    | 20 | 110            |

## 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 Флажол-Мартена. Оценивание моментов. Алгоритм Алон-Матияса-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.

## 4.3 Лабораторные работы

| № 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 распределителя. Группировка по ключу. Детали выполнения MapReduce. Комбинаторы. Обработка отказов узлов                                                                                                                     | 2            |
| 3    | 1         | Операции реляционной алгебры. Вычисление выборки с помощью MapReduce. Разбиение документов на шинглы. Хэширование документов с учетом близости. Методы для высокой степени сходства.                                               | 2            |
| 4    | 2         | Обработка потоков данных и специализированные алгоритмы для работы с данными.                                                                                                                                                      | 2            |
| 5    | 2         | Система управления потоками данных. Запросы к потокам. Проблемы обработки потоков. Выборка данных из потоков. Фильтрация потоков. Алгоритм Флажол-Мартена. Оценивание моментов. Алгоритм Алон-Матияса-Сегеди для вторых моментов.  | 2            |
| 6    | 2         | Получение репрезентативной выборки. Динамическое изменение размера выборки. Анализ фильтра Блума. Комбинирование оценок. Обработка бесконечных потоков.                                                                            | 2            |
| 7    | 3         | Алгоритмы кластеризации больших многомерных наборов данных. Иерархическая кластеризация. Альтернативные правила управления иерархической кластеризацией. Алгоритм к-средних                                                        | 2            |
| 8    | 3         | Алгоритмы машинного обучения, применяемые к большим наборам данных. Перцептроны. Механизм метода опорных векторов. Нахождение оптимальных приближений разделителей. Обучение по ближайшим соседям                                  | 2            |
| 9    | 3         | 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:                                                                                                                                                                                                                             | 20           |

## 5 Учебно-методическое обеспечение дисциплины

### 5.1 Основная литература

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 Нижегород. 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 информационные 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-Петербург, 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> (дата обращения: 02.06.2021). – 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. Лобачевского.- 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/](https://openedu.ru/course/hse/big_data_processing/) - «Открытое образование», Каталог курсов, MOOK: «Обработка больших данных»;

[https://openedu.ru/course/misis/ABD/?session=spring\\_2023](https://openedu.ru/course/misis/ABD/?session=spring_2023) - «Открытое образование», Каталог курсов, MOOK: «Введение в инженерии больших данных».

### 5.5 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. LibreOffice – свободно распространяемый офисный пакет программ, включающий в себя текстовый и табличный редакторы, редактор презентаций и другие офисные приложения.

3. Система управления учебным процессом Moodle, свободно распространяемая.

4. Программная система для организации видео-конференц-связи Webinar.ru, имеется лицензия, входит в реестр отечественного ПО.

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.