

Документ подписан простой электронной подписью
 Информация о владельце:
 ФИО: Мочалин Константин Сергеевич
 Должность: И.о. ректора
 Дата подписания: 02.07.2026 14:25:46
 Уникальный программный ключ:
 b7695d6b97247fced4385685adb0d9f8e6f2cdf

ФЕДЕРАЛЬНОЕ АГЕНТСТВО МОРСКОГО И РЕЧНОГО ТРАНСПОРТА

Федеральное государственное бюджетное образовательное учреждение высшего образования "Сибирский государственный университет водного транспорта"

Б1.О.21

Большие данные

рабочая программа дисциплины (модуля)

Закреплена за кафедрой	Информационных систем		
Образовательная программа	09.03.02 Направление подготовки "Информационные системы и технологии" Профиль "Проектирование информационных систем и их компонентов" год начала подготовки 2023		
Квалификация	бакалавр		
Форма обучения	очная		
Общая трудоемкость	3 ЗЕТ		
Часов по учебному плану	108	Виды контроля на курсах: зачет с оценкой 8	
в том числе:			
аудиторные занятия	36		
самостоятельная работа	68		

Распределение часов дисциплины по семестрам

Семестр (<Курс>.<Семестр на курсе>)	8 (4.2)		Итого	
Неделя	12 4/6			
Вид занятий	УП	РП	УП	РП
Лекции	24	24	24	24
Лабораторные	12	12	12	12
Иная контактная работа	4	4	4	4
Итого ауд.	36	36	36	36
Контактная работа	40	40	40	40
Сам. работа	68	68	68	68
Итого	108	108	108	108

Рабочая программа дисциплины

разработана в соответствии с ФГОС:

Федеральный государственный образовательный стандарт высшего образования - бакалавриат по направлению подготовки 09.03.02 Информационные системы и технологии (приказ Минобрнауки России от 19.09.2017 г. № 926)

составлена на основании учебного плана образовательной программы:

09.03.02 Направление подготовки "Информационные системы и технологии"
Профиль "Проектирование информационных систем и их компонентов"
год начала подготовки 2023

Рабочую программу составил(и):

к.т.н., Доцент, Гольшиев Д.Н.

Рабочая программа одобрена на заседании кафедры

Заведующий кафедрой Моторин Сергей Викторович

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1	Целью изучения дисциплины является обеспечение базового уровня подготовки обучающихся в области анализа и обработки больших данных.
-----	---

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ООП

Цикл (раздел) ООП:	Б1.О
2.1	Требования к предварительной подготовке обучающегося:
2.1.1	Моделирование систем
2.1.2	Научно-исследовательские проекты
2.1.3	Администрирование информационных систем
2.1.4	Технологии коммутации компьютерных сетей
2.1.5	Инструментальные средства информационных систем
2.1.6	Основы имитационного моделирования
2.1.7	Ситуационное моделирование информационных систем
2.1.8	Управление социально-трудовыми отношениями
2.1.9	Web-технологии и стандарты
2.1.10	Архитектура ЭВМ
2.1.11	Инфокоммуникационные системы и сети
2.1.12	Научно-исследовательская работа (получение первичных навыков научно-исследовательской работы)
2.1.13	Управление качеством
2.1.14	Моделирование систем
2.1.15	Научно-исследовательские проекты
2.1.16	Администрирование информационных систем
2.1.17	Технологии коммутации компьютерных сетей
2.1.18	Инструментальные средства информационных систем
2.1.19	Основы имитационного моделирования
2.1.20	Ситуационное моделирование информационных систем
2.1.21	Управление социально-трудовыми отношениями
2.1.22	Web-технологии и стандарты
2.1.23	Архитектура ЭВМ
2.1.24	Инфокоммуникационные системы и сети
2.1.25	Научно-исследовательская работа (получение первичных навыков научно-исследовательской работы)
2.1.26	Управление качеством
2.2	Дисциплины и практики, для которых освоение данной дисциплины (модуля) необходимо как предшествующее:

3. КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ (МОДУЛЯ)

УК-2: Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений

УК-2.1: Способен к целеполаганию и ранжированию задач в рамках поставленной цели

УК-2.2: Определяет оптимальные способы решения задач исходя из действующих правовых норм, имеющихся условий, ресурсов и ограничений

УК-2.3: Находит оптимальные способы решения задач исходя из имеющихся условий, ресурсов и ограничений

ОПК-2: Способен понимать принципы работы современных информационных технологий и программных средств, в том числе отечественного производства, и использовать их при решении задач профессиональной деятельности;

ОПК-2.1: Использует современные информационные технологии при решении задач профессиональной деятельности

ОПК-2.2: Использует программные средства, в том числе отечественного производства при решении задач профессиональной деятельности

ОПК-2.3: Использует методы моделирования (математического, графического, компьютерного) при решении задач профессиональной деятельности

ПК-1: Способен к выполнению работ и управлению работами по созданию (модификации) и сопровождению ИС, автоматизирующих задачи организационного управления и бизнес-процессы

ПК-1.7: Разрабатывает базы данных ИС

В результате освоения дисциплины обучающийся должен

3.1	Знать:
3.1.1	Основные методы анализа и обработки данных
3.2	Уметь:
3.2.1	Строить модели машинного обучения, работающие с большими данными
3.3	Владеть:
3.3.1	Инструментарием для анализа и обработки больших данных

4. СТРУКТУРА ДИСЦИПЛИНЫ (МОДУЛЯ)

Вид занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Литература	ПрПо дгот
Раздел	Раздел 1. Теория и практика больших данных				
Лек	Данные и метаданные /Лек/	8	4	Л1.1 Л1.2 Л1.3Л2.1	0
Лаб	Данные и метаданные /Лаб/	8	2	Л1.1 Л1.2 Л1.3Л2.1	0
Ср	Данные и метаданные /Ср/	8	10	Л1.1 Л1.2 Л1.3Л2.1	0
Лек	Большие данные /Лек/	8	4	Л1.1 Л1.2 Л1.3Л2.1	0
Ср	Большие данные /Ср/	8	4	Л1.1 Л1.2 Л1.3Л2.1	0
Лек	Технологии обработки больших данных /Лек/	8	8	Л1.1 Л1.2 Л1.3Л2.1	0
Лаб	Технологии обработки больших данных /Лаб/	8	4	Л1.1 Л1.2 Л1.3Л2.1	0
Ср	Технологии обработки больших данных /Ср/	8	12	Л1.1 Л1.2 Л1.3Л2.1	0
Лек	Методы анализа больших данных /Лек/	8	8	Л1.1 Л1.2 Л1.3Л2.1	0
Лаб	Методы анализа больших данных /Лаб/	8	6	Л1.1 Л1.2 Л1.3Л2.1	0
Ср	Методы анализа больших данных /Ср/	8	42	Л1.1 Л1.2 Л1.3Л2.1	0
ИКР	/ИКР/	8	4		0

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Раздел 1: «Теория и практика больших данных»

Тема 1 Данные и метаданные

Определения данных. Жизненный цикл данных. Создание данных: приобретение, запись, регистрация сигналов.

Обслуживание данных: перемещение, интеграция, очистка, обогащение, ETL процедуры. Синтез данных. Использование данных. Публикация данных. Архивация данных. Уничтожение данных.

Типы данных: мастер-данные, транзакционные, ретроспективные, ссылочные. Форматы данных: структурированные, полуструктурированные, неструктурированные.

Понятие метаданных. Классификация метаданных. Форматы метаданных. Схемы метаданных (информационные онтологии).

Тема 2 Большие данные

Понятие больших данных. Основные характеристики больших данных. Набор признаков VVV. Источники больших данных: Интернет вещей, социальные медиа, геоданные, наука, медицина, производство, экономика, корпоративные данные. Классификация больших данных.

Тема 3 Технологии обработки больших данных

Вертикальная и горизонтальная масштабируемость. Параллельная, распределённая обработка данных. Требования к транзакционной системе ACID, BASE. Теорема CAP (Брюера). Краудсорсинг.

Технология NoSQL. Модель «ключ-значение». Модель семейства столбцов. Документная БД. Графовая БД. Технология MapReduce. Функции Map, Reduce. Технология Hadoop. Модули Common, HDFS, YARN, MapReduce. Технология R.

Синтаксис языка, графический интерфейс.

Аппаратные решения для обработки больших данных.

Тема 4 Методы анализа больших данных

Методы класса Data Mining. Машинное обучение. Искусственные нейронные сети, генетические алгоритмы. Цифровая обработка сигналов. Предсказательная аналитика. Имитационное моделирование. Распознавание образов.

Пространственный анализ. Статистический анализ. Анализ временных рядов. A/B-тестирование. Визуализация данных.

Содержание лабораторных работ:

Создание, обслуживание, синтез данных. (2 часа)

Метаданные. Информационные онтологии. (2 часа)

Технологии обработки больших данных. (4 часа)

Методы класса Data Mining. (4 часов)

6. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

6.1. Перечень видов оценочных средств

Вопросы к зачету с оценкой

Вопросы к лабораторным работам

6.2. Темы письменных работ

Не предусмотрен

6.3. Контрольные вопросы и задания

ТИПОВЫЕ ТЕСТОВЫЕ ЗАДАНИЯ по теме «Большие данные (Big Data)». Задания охватывают ключевые аспекты: определение, характеристики, технологии, применение, этика и вызовы.

Блок 1. Основы Big Data

1. Что означает термин «Big Data»?

- а) Любые данные, хранящиеся в облаке
- б) Совокупность технологий для обработки малых объёмов данных
- в) Данные, которые невозможно обработать стандартными средствами из-за их объёма, скорости поступления и/или неоднородности
- г) Только текстовые данные из социальных сетей

Ответ: в

2. Какие три «V» традиционно характеризуют Big Data?

- а) Volume, Velocity, Variety
- б) Value, Veracity, Volume
- в) Velocity, Variability, Validity
- г) Volume, Variety, Valuation

Ответ: а

3. Что добавляет четвёртое «V» (Veracity) к классической триаде Big Data?

- а) Скорость обработки
- б) Достоверность и качество данных
- в) Объём данных
- г) Разнообразие форматов

Ответ: б

4. Какой примерный порог объёма данных считается «большим» в современных системах?

- а) 1 ГБ
- б) 1 ТБ
- в) 1 ПБ (петабайт)
- г) Нет универсального порога — зависит от контекста

Ответ: г

5. Что такое «Data Lake»?

- а) База данных с жёсткой схемой (schema-on-write)
- б) Хранилище необработанных данных в исходном формате
- в) Инструмент для визуализации данных
- г) Протокол передачи данных

Ответ: б

Блок 2. Технологии и инструменты

6. Какая система используется для распределённого хранения Big Data?

- а) MySQL
- б) Hadoop HDFS
- в) Microsoft Access
- г) SQLite

Ответ: б

7. Что такое MapReduce?

- а) Язык запросов для SQL
- б) Модель параллельных вычислений для обработки больших наборов данных
- в) Тип базы данных
- г) Алгоритм машинного обучения

Ответ: б

8. Какой инструмент чаще всего используется для интерактивной аналитики Big Data?

- а) Excel
- б) Apache Spark
- в) Notepad++
- г) PowerPoint

Ответ: б

9. Для чего применяется Apache Kafka?

- а) Для хранения структурированных данных
- б) Для потоковой передачи и обработки данных в реальном времени
- в) Для визуализации графиков
- г) Для шифрования данных

Ответ: б

10. Что такое NoSQL?

- а) Улучшенная версия SQL
- б) Класс СУБД, не использующих SQL и предназначенных для масштабируемых данных
- в) Язык программирования
- г) Протокол безопасности

Ответ: б

Блок 3. Обработка и анализ

11. Какой метод используется для снижения размерности данных?

- а) Линейная регрессия
- б) PCA (Principal Component Analysis)
- в) Деревья решений
- г) K-means

Ответ: б

12. Что такое ETL в контексте Big Data?

- а) Extract, Transform, Load — процесс извлечения, преобразования и загрузки данных
- б) Encrypt, Transfer, Log — шифрование, передача, логирование
- в) Evaluate, Test, Launch — оценка, тестирование, запуск
- г) None of the above

Ответ: а

13. Какой алгоритм часто используется для кластеризации больших данных?

- а) SVM
- б) K-means
- в) Logistic Regression
- г) Random Forest

Ответ: б

14. Что означает «real-time analytics»?

- а) Анализ данных раз в сутки
- б) Анализ данных с задержкой более 1 часа
- в) Анализ данных в момент их поступления
- г) Анализ исторических данных

Ответ: в

15. Какой инструмент используется для визуализации больших данных?

- а) Tableau
- б) Word
- в) Paint
- г) CMD

Ответ: а

Блок 4. Применение и кейсы

16. В какой отрасли Big Data применяется для предсказания оттока клиентов (churn prediction)?

- а) Телеком
- б) Сельское хозяйство
- в) Библиотечное дело
- г) Полиграфия

Ответ: а

17. Как Big Data используется в здравоохранении?

- а) Для анализа геномов и персонализированной медицины
- б) Для печати медицинских карт
- в) Для отправки SMS-уведомлений
- г) Для учёта канцелярских принадлежностей

Ответ: а

18. Какой кейс относится к «умному городу» (Smart City)?

- а) Анализ трафика для оптимизации светофоров
- б) Ручная уборка улиц
- в) Печать афиш
- г) Опросы жителей

Ответ: а

19. Как Big Data помогает в ритейле?

- а) Для прогнозирования спроса и персонализации предложений
- б) Для ручной инвентаризации
- в) Для оформления витрин
- г) Для найма персонала

Ответ: а

20. Какой пример относится к «интернету вещей» (IoT) и Big Data?

- а) Данные с датчиков умного дома
- б) Письмо по электронной почте
- в) Книга в библиотеке
- г) Телефонный звонок

Ответ: а

Блок 5. Этика, безопасность, вызовы

21. Какая проблема связана с конфиденциальностью в Big Data?

- а) Избыточная яркость экранов
- б) Утечка персональных данных
- в) Медленная загрузка сайтов
- г) Шум в офисах

Ответ: б

22. Что такое «data bias»?

- а) Ошибка в коде
- б) Предвзятость данных, ведущая к несправедливым выводам
- в) Потеря данных
- г) Дублирование записей

Ответ: б

23. Какой стандарт регулирует защиту персональных данных в ЕС?

- а) HIPAA
- б) GDPR
- в) ISO 9001
- г) PCI DSS

Ответ: б

24. Что такое «data governance»?

- а) Управление политиками, процессами и качеством данных
- б) Хранение данных в облаке
- в) Удаление данных
- г) Передача данных по сети

Ответ: а

25. Какой риск связан с хранением Big Data?

- а) Увеличение числа сотрудников
- б) Рост затрат на инфраструктуру и угрозы кибератак
- в) Уменьшение объёма данных
- г) Снижение скорости интернета

Ответ: б

Блок 6. Перспективы и тренды

26. Что такое «Data Science»?

- а) Наука о хранении данных
- б) Междисциплинарная область, объединяющая статистику, машинное обучение и анализ данных
- в) Раздел физики
- г) Тип базы данных

Ответ: б

27. Какой тренд связан с «edge computing» в Big Data?

- а) Обработка данных ближе к источнику (на устройствах)
 б) Централизованное хранение в одном дата-центре
 в) Отказ от облачных технологий
 г) Использование бумажных носителей

Ответ: а

28. Что означает «AI-driven analytics»?

- а) Анализ данных вручную
 б) Использование ИИ для автоматического анализа и прогнозов
 в) Печать отчётов
 г) Отправка данных по почте

Ответ: б

29. Какой фактор критичен для успеха Big Data-проектов?

- а) Наличие качественных данных и чётких бизнес-целей
 б) Покупка самого дорогого оборудования
 в) Игнорирование этики
 г) Отказ от визуализации

Ответ: а

30. Что такое «data monetization»?

- а) Превращение данных в источник дохода (например, через аналитику или продажу инсайтов)
 б) Печать денег
 в) Обмен данными без цели
 г) Удаление данных

Ответ: а

Блок 7. Принципы и вызовы (открытые вопросы)

31. Объясните, почему «достоверность» (Veracity) добавляют как 4 ю характеристику Big Data (4V).

Приведите пример проблемы с Veracity.

32. Опишите один сценарий использования Big Data в здравоохранении. Укажите:

- Какие данные анализируются?
- Какую проблему это решает?
- Какой инструмент/технологию можно применить?

33. Перечислите 3 основных вызова при работе с Big Data и предложите по одному решению для каждого.

Ответы:

31. Veracity отражает качество данных (шум, противоречия). Пример: фейковые отзывы в интернет-магазинах.

32. Пример: анализ электронных медкарт для выявления рисков заболеваний. Данные: анамнез, анализы.

Инструмент: Hadoop + ML-алгоритмы.

33. Вызовы: хранение (решение — облачные хранилища), обработка (Spark), конфиденциальность (шифрование).

Блок 8. Анализ кейсов (развёрнутые ответы)

34. Кейс: Розничная сеть хочет прогнозировать спрос на товары.

- Какие источники данных ей понадобятся?
- Какую модель машинного обучения можно использовать?
- Как оценить качество прогноза?

35. Кейс: Соцсеть анализирует поведение пользователей для таргетированной рекламы.

- Какие типы данных она будет собирать?
- Какие риски конфиденциальности возникают?
- Как их минимизировать?

Ответы:

34. Источники: продажи, погода, соцсети. Модель: ARIMA или нейросети. Качество: MAE, RMSE.

35. Данные: лайки, время в приложении. Риски: утечка персональных данных. Минимизация: анонимизация, GDPR-комплаенс.

Блок 5. Терминология (сопоставление)

36. Сопоставьте термин и его описание:

Термин

1. ETL
2. Data Lake
3. NoSQL
4. Stream Processing

Описание

- А. Система для обработки потоков данных в реальном времени
 Б. Процесс извлечения, преобразования и загрузки данных
 В. Хранилище для сырых данных в любом формате
 Г. База данных без жёсткой схемы, оптимизированная для Big Data

Ответ: 1 Б, 2 В, 3 Г, 4 А.

Блок 6. Расчёты и логика

37. Сопоставьте термин и его описание:
- Объем данных: 10 ТБ
 - Скорость записи: 100 МБ/с
 - Сколько времени потребуется для записи всего объема? Ответ дайте в часах.
38. В наборе данных 1 млн записей. 5% содержат ошибки. Сколько записей нужно очистить?
- Ответы:
37. $\approx 27,8$ часов.
38. 50 тыс. записей.

6.4. Методические материалы, определяющие процедуры оценивания

Итоговая оценка является арифметической суммой всех баллов, полученных обучающимся в процессе изучения дисциплины. В учёт итоговой оценки по данной методике принимается шкала оценивания каждого вида занятий по дисциплине: лекции, практики, лабораторные работы, семинары и т.д. Преподавателем на первом занятии озвучиваются максимальное количество баллов, которое можно получить за данный вид занятий. Вес каждого вида занятий в баллах зависит от их объема и утверждается на первом заседании кафедры в текущем учебном году. Балльная шкала по видам занятий для дисциплины приведена ниже.

Методика получения итоговой оценки по 4-х балльной шкале:

5 (отлично)	≥ 85
4 (хорошо)	$75 \div 84$
3 (удовлетворительно)	$51 \div 74$
2 (неудовлетворительно)	≤ 50

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

7.1 Рекомендуемая литература

7.1.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год
Л1.1	Парфенов Ю. П., Папуловская Н. В.	Постреляционные хранилища данных: Учебное пособие	Москва: Издательство Юрайт, 2018
Л1.2	Мхитарян В. С., Архипова М. Ю., Дуброва Т. А., Миронкина Ю. Н., Сиротин В. П.	Анализ данных: Учебник	Москва: Издательство Юрайт, 2019
Л1.3	Станкевич Л. А.	Интеллектуальные системы и технологии: Учебник и практикум	Москва: Издательство Юрайт, 2019

7.1.2. Дополнительная литература

	Авторы, составители	Заглавие	Издательство, год
Л2.1	Бессмертный И. А., Нугуманова А. Б., Платонов А. В.	Интеллектуальные системы: Учебник и практикум	Москва: Издательство Юрайт, 2018

7.2 Перечень ресурсов информационно-телекоммуникационной сети "Интернет"

Э1	Open Data Science. Крупнейшее русскоязычное Data Science сообщество
Э2	Школа Больших Данных

8. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Назначение	Оборудование
Учебная аудитория для проведения занятий лекционного типа	Аудиторная доска; Комплект учебной мебели; Мультимедийное оборудование: проектор (стационарный), Экран (стационарный), ПК (стационарный)
Учебная аудитория для проведения занятий лекционного типа	Аудиторная доска; Комплект учебной мебели; Мультимедийное оборудование: проектор (стационарный), экран (стационарный), ПК (стационарный)
Компьютерный класс - Лаборатория информационных систем - учебная аудитория для проведения лабораторных занятий	Аудиторная доска; Комплект учебной мебели; ПК-9 шт. (в т.ч. преподавательский); Мультимедийное оборудование: проектор, экран, ПК (переносной)
Компьютерный класс - Лаборатория информационных систем - учебная аудитория для проведения лабораторных занятий	Аудиторная доска; Комплект учебной мебели; ПК-9 шт. (в т.ч. преподавательский); Мультимедийное оборудование: проектор, экран, ПК (переносной)

занятий		
Помещение самостоятельной обучающихся	для работы	Комплект учебной мебели; ПК – 6 шт., подключенных к сети "Интернет" и обеспечивающих доступ в электронную информационно-образовательную среду Университета.