МИНОБРНАУКИ РОССИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»
(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ
Заведующий кафедрой
Математических методов исследования операций
Азарнова Т.В.
22.03.2024 г

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.В.ДВ.03.02.04 Анализ больших данных

1. Код и наименование направления подготовки / специальности:

01.03.02 Прикладная математика и информатика

2. Профиль подготовки / специализация/магистерская программа:

Прикладная математика и компьютерные технологии

3. Квалификация (степень) выпускника: бакалавр

4. Форма обучения: очная

5. Кафедра, отвечающая за реализацию дисциплины: ММИО

6. Составители программы: Ухлова В.В., к.ф.-м.н, доцент кафедры ММИО

7. Рекомендована: НМС факультета Прикладной математики, информатики и механики № 5 от 22.03.2024

8. Учебный год: 2027/2028 **Семестр(ы):** 8

9. Цели и задачи учебной дисциплины

Целями освоения учебной дисциплины являются: формирование целостного представления о современных технология работы с данными, методам и алгоритмам работы с большими массивами данных, которые позволяют обрабатывать, интерпретировать, оформлять и представлять профессиональному обществу результаты исследований.

Задачи учебной дисциплины:

– изучение процессов консолидации, анализа, обработки больших данных;

– получение знаний и умений, необходимых для проведения анализа предметной области, выявления информационных потребностей организации;

– получение знаний и умений, необходимых для проведения аналитического исследования в соответствии с согласованными требованиями;

– приобретение навыков формализации требований к данным, формализации и построения концептуальной модели решения профессиональных задач;

– приобретение навыков разработки требований и выбора инструментальных средств и технологий проектирования прикладных моделей и программ.

10. Место учебной дисциплины в структуре ОПОП:

дисциплина относится к части, формируемой участниками образовательных отношений, блока Б1 дисциплин учебного плана.

11. Планируемые результаты обучения по дисциплине/модулю (знания, умения, навыки), соотнесенные с планируемыми результатами освоения образовательной программы (компетенциями) и индикаторами их достижения

Код	Название компетенции	Код(ы)	Индикаторы(ы)	Планируемые результаты обучения
ПК-4	Способен разрабатывать комплекс требований к программному обеспечению, осуществлять его проектирование с учетом особенностей предметной области для решения прикладных задач в естественных науках, промышленности и бизнесе и управлять работами по созданию (модификации) и сопровождению информационных ресурсов	ПК-4.1	Проводит анализ и формализацию предметной области, выявляет информационные потребности и оценивает возможности реализации требований к компьютерному программному обеспечению и ИР.	Знать: основные технологии консолидации, обработки и управления большими данными, позволяющие осуществлять поиск, сбор и хранение информации из открытых источников и специализированных баз данных; основные методологии анализа данных; алгоритмы обработки данных. основные методики исследования и испытания разработанных методов, моделей, алгоритмов, технологий и инструментальных средств по работе с данными. Уметь: осуществлять информационный поиск с использованием открытых источников информации и специализированных баз данных; использовать инструментальные средства для работы с данными, в том числе, с большими данными; проводить исследования и испытания методов, моделей, алгоритмов и
ПК-6	Способен проводить обработку и анализ больших данных с использованием существующей в	ПК-6.1	Выявляет, формирует и согласовывает требования к результатам аналитических работ, в том числе, с использованием технологий больших данных.	

	организации методологической и технологической инфраструктуры	ПК-6.2	Осуществляет планирование, организацию и подготовку данных для проведения аналитических работ, в том числе, с использованием технологий больших данных, а также осуществляет выполнение указанных работ.	инструментальных средств работы с большими данными. Владеть навыками инсталляции и настройки ПО для работы с большими данными.
		ПК-6.3	Проводит аналитическое исследование в соответствии с согласованными требованиями заказчика, в том числе, с использованием технологий больших данных.	

12. Объем дисциплины в зачетных единицах/часах в соответствии с учебным планом —3/108

Форма промежуточной аттестации – зачет с оценкой.

13. Трудоемкость по видам учебной работы

Вид учебной работы	Трудоемкость (часы)				
	Всего	В том числе в интерактивной форме	По семестрам		
			№ сем. 8	№ сем.	…..
Аудиторные занятия					
в том числе: лекции	32		32		
практические	-		-		
лабораторные	16		16		
Самостоятельная работа	60		60		
Форма промежуточной аттестации	Зачет		Зачет		
Итого:	108		108		

13.1. Содержание дисциплины

№ п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
		1. Лекции	
1.1	Понятие Data science и Big Data, область применения, тенденции развития	1.1.1Термины и определения. Особенности технологий. Сферы применения, состояние и тенденции развития. 1.1.2 Бизнес-кейсы Big Data. 1.1.3 Рынок Big Data в России и мире.	Анализ больших данных (01.03.02, Ухлова В.В.)
1.2	Технологии консолидации, обработки и управления большими данными	1.2.1 Платформа Hadoop: архитектура и принцип работы. Организация файловой системы HDFS. Концепция Map Reduce. Система YARN. Экосистема Hadoop.	
		1.2.2 Платформа Spark: архитектура и принцип работы. Файловые системы для работы в Spark.	

		1.2.3 Сравнение Hadoop и Spark: инфраструктура, работа ПО, задачи.	
		1.2.4 Базы данных NoSQL	
1.3	Основные процессы в Data science	1.3.1. Жизненный цикл аналитики больших данных: процессы сбора, подготовки, исследования и отображения данных. 1.3.2 Методы моделирования данных. 1.3.3 Визуализация данных.	
		2. ***Лабораторные работы***	
2.1	Методы работы с данными	2.1.1 Загрузка данных. Проверка качества данных. Очистка данных. Отображение данных. 2.1.2 Организация хранения данных. 2.1.3 Методы обработки и анализа данных. 2.1.4 Инструменты управления данными. 2.1.5 Выбор и установка ПО для работы с большими данными.	Анализ больших данных (01.03.02, Ухлова В.В.)

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				
		Лекции	Практи-ческие	Лабораторные	Самостоятельная работа	Всего
1	Понятие Data science и Big Data, область применения	4	-	-	6	10
2	Основные процессы в Data science	10	-	-	10	20
3	Технологии консолидации, обработки и управления большими данными	16	-	4	14	34
4	Методы работы с данными	2	-	12	30	44
	Итого:	32	-	16	60	108

14. Методические указания для обучающихся по освоению дисциплины

Дисциплина реализуется по тематическому принципу, каждая тема представляет собой завершенный раздел курса. Темы с кодировкой Х.Х.1 относятся к базовому (обязательному) блоку для обучения. На первом занятии студент получает информацию для доступа к комплексу учебно-методических материалов.

Лекционные занятия посвящены рассмотрению теоретических основ дисциплины: вводятся основные понятия, изучаются базовые технологии, разбираются основные процессы работы с большими данными. Лабораторные работы предназначены для формирования умений и навыков, закрепленных компетенций по ОПОП. Они организовываются в виде выполнения отдельных заданий. По окончании изучения дисциплины проводится тестирование.

Самостоятельная работа студентов включает в себя проработку учебного материала лекций, разбор заданий лабораторных работ, подготовку к экзамену. Для успешного освоения дисциплины рекомендуется подробно конспектировать лекционный материал, просматривать презентации по соответствующей теме, чтобы систематизировать изучаемый материал, выполнять задания лабораторных работ.

Промежуточная аттестация по результатам обучения проводится в форме экзамена, контролирующего освоение ключевых положений дисциплины, составляющих основу знаний по дисциплине.

При использовании дистанционных образовательных технологий и электронного обучения следует выполнять все указания преподавателя по работе на LMS-платформе, своевременно подключаться к online-занятиям, соблюдать рекомендации по организации самостоятельной работы.

15. Перечень основной и дополнительной литературы, ресурсов интернет, необходимых для освоения дисциплины

а) основная литература:

№ п/п	Источник
1	Основы технологий Big Data [Электронный ресурс] : учебное пособие / Воронеж. гос. ун-т / В.В. Ухлова .— Электрон. текстовые дан. — Воронеж : Издательский дом ВГУ, 2020 .— Загл. с титула экрана .— Свободный доступ из интрасети ВГУ .— Текстовый файл .— <URL:http://www.lib.vsu.ru/ >.
2	Макшанов, А. В. Большие данные. Big Data / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 2-е изд., стер. — Санкт-Петербург : Лань, 2022. — 188 с. — ISBN 978-5-8114-9690-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/198599 (дата обращения: 25.02.2024). — Режим доступа: для авториз. пользователей.
3	Макшанов, А. В. Современные технологии интеллектуального анализа данных : учебное пособие для спо / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — Санкт-Петербург : Лань, 2020. — 228 с. — ISBN 978-5-8114-5451-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/149343 (дата обращения: 25.02.2024). — Режим доступа: для авториз. пользователей.

б) дополнительная литература:

№ п/п	Источник
4	Макшанов, А. В. Системы поддержки принятия решений : учебное пособие для вузов / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — 2-е изд., стер. — Санкт-Петербург : Лань, 2021. — 108 с. — ISBN 978-5-8114-8489-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/176903 (дата обращения: 25.02.2024). — Режим доступа: для авториз. пользователей.
5	Юре, Л. Анализ больших наборов данных / Л. Юре, Р. Ананд, Д. У. Джеффри ; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2016. — 498 с. — ISBN 978-5-97060-190-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/93571 (дата обращения: 25.02.2024). — Режим доступа: для авториз. пользователей.

в) информационные электронно-образовательные ресурсы (официальные ресурсы интернет)*:

№ п/п	Ресурс
6	Электронно-библиотечная система «Лань» - Режим доступа: https://e.lanbook.com
7	Электронный каталог Научной библиотеки Воронежского государственного университета. – Режим доступа: http//:www.lib.vsu.ru.
8	Анализ больших данных (01.03.02, Ухлова В.В.)/ В.В. Ухлова. — Образовательный портал «Электронный университет ВГУ». — Режим доступа: https://edu.vsu.ru/course/view.php?id=5525

16. Перечень учебно-методического обеспечения для самостоятельной работы

Самостоятельная работа обучающегося должна включать подготовку к тестированию, лабораторным занятиям и подготовку к промежуточной аттестации. При самостоятельной подготовку обучающийся пользуется конспектами лекций и литературой по тематике лекционного материала, заданий контрольных и лабораторных работ. Для обеспечения самостоятельной работы студентов в электронном курсе дисциплины на образовательном портале «Электронный университет ВГУ» сформирован учебно-методический комплекс, который включает в себя: программу курса, учебные пособия и справочные материалы, методические

указания по выполнению лабораторных работ. Студенты получают доступ к данным материалам на первом занятии по дисциплине.

17. Образовательные технологии, используемые при реализации учебной дисциплины, включая дистанционные образовательные технологии (ДОТ), электронное обучение (ЭО), смешанное обучение):

При реализации дисциплины используются следующие образовательные технологии: логическое построение дисциплины, обозначение теоретического и практического компонентов в учебном материале. Применяются разные типы лекций (вводная, обзорная, информационная, проблемная). Дисциплина реализуется с применением информационно-коммуникационных технологий.

Информационно-коммуникативные технологии для реализации учебной дисциплины:
- технологии синхронного и асинхронного взаимодействия студентов и преподавателя посредством служб (сервисов) по пересылке и получению электронных сообщений, в том числе, по сети Интернет;
- сервис электронной почты для оперативной связи преподавателя и студентов.

Дисциплина реализуется с применением электронного обучения и дистанционных образовательных технологий, для организации самостоятельной работы обучающихся используется онлайн-курс, размещенный на платформе Электронного университета ВГУ (LMS moodle), а также другие Интернет-ресурсы, приведенные в п.15в.

18. Материально-техническое обеспечение дисциплины:

Лекционная аудитория должна быть оборудована учебной мебелью, компьютером, мультимедийным оборудованием (проектор, экран, средства звуковоспроизведения), допускается переносное оборудование.

Лабораторные работы должны проводиться в специализированной аудитории, оснащенной учебной мебелью и персональными компьютерами с доступом в сеть Интернет (компьютерные классы, студии), мультимедийным оборудованием (проектор, экран, средства звуковоспроизведения), Число рабочих мест в аудитории должно быть таким, чтобы обеспечивалась индивидуальная работа студента персональном компьютере.

Для самостоятельной работы необходимы компьютерные классы, помещения, оснащенные компьютерами с доступом к сети Интернет и платформе Электронного университета ВГУ (LMS moodle).

Программное обеспечение:
- ОС Windows 10, ОС Linux;
- пакет стандартных офисных приложений для работы с документами, таблицами и т.п. (МойОфис, LibreOffice);
- ПО Adobe Reader;
- специализированное ПО (ПО MatLab, Power BI);
- интернет-браузер (Mozilla Firefox, Яндекс).

19. Оценочные средства для проведения текущей и промежуточной аттестаций

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

- контрольная работа,

- тест,
- лабораторная работа.

Порядок оценки освоения обучающимися учебного материала определяется содержанием следующих разделов дисциплины:

№ п/п	Наименования раздела дисциплины	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1	Понятие Data science и Big Data, область применения.	ПК-4	ПК-4.1	Контрольная работа
2	Технологии консолидации, обработки и управления большими данными.	ПК-6	ПК-6.1, ПК-6.2, ПК-6.3	Лабораторная работа 1-5
3	Основные процессы в Data science.	ПК-6	ПК-6.1, ПК-6.2, ПК-6.3	Лабораторная работа 1-5
4	Методы работы с данными	ПК-6	ПК-6.1, ПК-6.2, ПК-6.3	Лабораторная работа 1-5
Промежуточная аттестация, форма контроля – зачет с оценкой				Перечень вопросов

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

Контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

- контрольная работа,
- лабораторная работа.

Контрольная работа может быть заменена на тест, в зависимости от технологий обучения. При варианте ДО рекомендуется контрольную работу заменить на тест. Контрольная работа и тест являются взаимозаменяемыми.

Перечень заданий контрольной работы

Для исходного набора данных:

1) выполнить описание «идеальных» данных (тип данных, ограничения, шаблон и т.п);
2) привести варианты возможных ошибок в данных;
3) составить алгоритм повышения качества данных;
4) продемонстрировать траекторию изменения данных при использовании разработанного алгоритма;
5) составить рекомендации, позволяющие получать исходный набор данных с более высоким качеством.

Технология проведения

В качестве исходных данных студент берет любой набор из открытых источников (в формате xls/xlsx (количество записей должно быть более 50, атрибутов более 10). Если качество данных набора очень высокое, то искусственно «ухудшает» его.

Выполнение задания предусматривает использование информации из учебной и справочной литературы, а также ресурсов сети Интернет.

Технология проведения

В качестве исходных данных студент берет любой набор из открытых источников (в формате xls/xlsx (количество записей должно быть более 50, атрибутов более 10). Если качество данных набора очень высокое, то искусственно «ухудшает» его.

Выполнение задания предусматривает использование информации из учебной и справочной литературы, а также ресурсов сети Интернет.

Критерии оценивания:

- оценка «зачтено» выставляется студенту, если студент дал правильные ответы на 90 и более процентов заданий;
- оценка «не зачтено» - даны правильные ответы на менее чем 90 процентов заданий.

Перечень заданий теста

Пример компоновки вопросов теста (вопросы с вариантами ответов).
Вариант 1.

1. Приведите основные характеристики больших данных:

а) Virtualization, Volume, Variability, Vehicle;
б) Variety, Velocity, Volume, Value;
в) Verification, Volume, Velocity, Visualization;
г) Video, Value, Variety, Volume.

2. Расставьте в правильном порядке основные этапы процесса Data Science:

а) назначение цели исследования, сбор данных, подготовка данных, исследование данных, моделирование данных, отображение данных;
б) назначение цели исследования, сбор данных, подготовка данных, моделирование данных, исследование данных, отображение данных;
в) назначение цели исследования, подготовка данных, сбор данных, моделирование данных, исследование данных, отображение данных;
г) назначение цели исследования, сбор данных, подготовка данных, отображение данных, исследование данных, моделирование данных.

3.Поясните понятие:

Hadoop представляет собой…
а) набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах;
б) распределённую СУБД, позволяющую обрабатывать большие данные;
в) язык выполнения заданий в парадигме MapReduce;
г) распределённую файловую систему для организации хранения файлов большого объёма.

4. Принцип MapReduce состоит в том, чтобы
а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-параллельной обработки.

Технология проведения
Тест включает в себя 30 вопросов, вариант теста вбирается исходя из номера зачетки (последней цифры). Время на тестирование рассчитывается из соотношения 10 вопросов – 15 минут. Результаты тесты проверяются по ключу правильных ответов.

Критерии оценивания:
- оценка «зачтено» выставляется студенту, если студент дал правильные ответы на 90 и более процентов заданий (тест пройден);
- оценка «не зачтено» - даны правильные ответы на менее чем 90 процентов заданий (тест не пройден).

Перечень заданий для лабораторных работ.

Лабораторная работа №1
Пример задания.
Выполнить загрузку данных в аналитический контур. В качестве исходных данных использовать форматы xlsx, txt, pdf. Проверить факт загрузки с использованием инструментов отображения данных.

Технология проведения
Студент выбирает вариант задания, ориентируясь на номер зачетки (последняя цифра). Файлы исходных данных заранее должны быть размещены на сервере (компьютере студента). Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:
- оценка «зачтено» выставляется студенту, если задания выполнены в полном объеме;
- оценка «не зачтено» - работа не выполнена или выполнена не в полном объеме.

Лабораторная работа №2
Пример задания.
Вычислить основные статистики данных, загруженных в аналитический контур. В качестве исходных данных использовать форматы xlsx, txt, pdf. Отобразить полученные статистики с использованием соответствующих инструментов.

Технология проведения
Студент выбирает вариант задания, ориентируясь на номер зачетки (последняя цифра). Файлы исходных данных заранее должны быть размещены на сервере (компьютере студента). Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:
- оценка «зачтено» выставляется студенту, если задания выполнены в полном объеме;
- оценка «не зачтено» - работа не выполнена или выполнена не в полном объеме.

Лабораторная работа №3
Пример задания.
Выполнить проверку данных, загруженных в аналитический контур. В качестве исходных данных использовать форматы xlsx, txt, pdf. Повысить качество данных с использованием соответствующих инструментов.

Технология проведения

Студент выбирает вариант задания, ориентируясь на номер зачетки (последняя цифра). Файлы исходных данных заранее должны быть размещены на сервере (компьютере студента). Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:

- оценка «зачтено» выставляется студенту, если задания выполнены в полном объеме;
- оценка «не зачтено» - работа не выполнена или выполнена не в полном объеме.

Лабораторная работа №4

Пример задания.

Рассчитать метрики для данных, загруженных в аналитический контур. В качестве исходных данных использовать форматы xlsx, txt, pdf. Отобразить полученные метрики с использованием соответствующих инструментов.

Технология проведения

Студент выбирает вариант задания, ориентируясь на номер зачетки (последняя цифра). Файлы исходных данных заранее должны быть размещены на сервере (компьютере студента). Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:

- оценка «зачтено» выставляется студенту, если задания выполнены в полном объеме;
- оценка «не зачтено» - работа не выполнена или выполнена не в полном объеме.

Лабораторная работа №5

Пример задания.

Выполнить визуализацию обработанных данных, отобразить рассчитанные в предыдущих работах статистики и метрики. В качестве исходных данных использовать форматы xlsx, txt, pdf.

Технология проведения

Студент выбирает вариант задания, ориентируясь на номер зачетки (последняя цифра). Файлы исходных данных заранее должны быть размещены на сервере (компьютере студента). Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:

- оценка «зачтено» выставляется студенту, если задания выполнены в полном объеме;
- оценка «не зачтено» - работа не выполнена или выполнена не в полном объеме.

Лабораторная работа №6

Пример задания.

Выполнить расчет хранилища данных для системы офисной системы видеонаблюдения. Параметры системы видеонаблюдения: 5 камер, разрешение 2.1, 1920x1080, частота 12к/с, кодек H.264.Период хранения данных составляет 3 месяца,

Технология проведения

Студент выбирает вариант задания, ориентируясь на номер зачетки (последняя цифра). Время выполнения задания составляет 3 часа. Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:

- оценка «отлично» выставляется студенту, если работа выполнена в полном объеме (приведены все расчеты и они правильные, даны пояснения);
- оценка «хорошо» - работа выполнена полностью, но имеются незначительные ошибки;

- оценка «удовлетворительно» - работа выполнена полностью, но в представленной части много ошибок или представлена часть работы и она без ошибок;
- оценка «неудовлетворительно» - работа не выполнена.

Лабораторная работа №7
Пример задания.

1. Обозначить бизнес-проблему.
2. Сформулировать бизнес-цели.
3. Обозначить бизнес-задачи.
4. Свести бизнес-задачу к аналитической задаче.
5. Определить потребности в ресурсах (указать источники, виды ресурсов, виды и содержание информации, которую можно получить).
6. Подобрать технологии (методы, модели, алгоритмы, инструментальные средства), позволяющие работать с определенными в п.6 ресурсами.
7. При необходимости дать рекомендации по доработке технологии (методы, модели, алгоритмы, инструментальные средства) из п.6.

Технология проведения
Предметную область студент выбирает самостоятельно, базируясь на информации из открытых источников. Время выполнения задания составляет 3 часа. Студенту разрешается пользоваться информацией из открытых источников.

Критерии оценивания:

- оценка «отлично» выставляется студенту, если работа выполнена в полном объеме, полученные результаты аргументированы;
- оценка «хорошо» - работа выполнена полностью, но полученные результаты не логичны или требуют уточнения;
- оценка «удовлетворительно» - работа выполнена полностью, но имеет место большое количество ошибок или представлена часть работы и она без ошибок;
- оценка «неудовлетворительно» - работа не выполнена.

Лабораторные работы №6-7 могут заменяться на подготовку реферата (презентацию) по одному из разделов дисциплины (одна работа равна одному реферату).

Примерные темы рефератов (презентаций)

Обзор инструментов работы с данными
Пример подбора инструментов обработки данных
Пример подбора инструментов очистки данных
Пример подбора инструментов отображения данных
Пример расчета статистик загруженных данных в ПО
Пример расчета метрик загруженных данных в ПО
Разработка алгоритма анализа больших данных в рамках поставленной задачи

Технология проведения
Тема выбирается обучающимся самостоятельно. При этом рекомендуется выбор тем в группе таким образом, чтобы они не повторялись.

Критерии оценки:

- оценка «зачтено» выставляется студенту, если:
  - изложенная информация является актуальной на момент представления реферата;
  - по содержанию реферат отражает все основные аспекты выбранной темы;
  - реферат оформлен в соответствии с рекомендациями по оформлению;
- оценка «не зачтено», если:
  - изложенная информация не является актуальной на момент представления реферата;
  - по содержанию реферат не отражает все основные аспекты выбранной темы;
  - реферат не оформлен в соответствии с рекомендациями по оформлению.

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: тест.

Тестовые задания.

Пример компоновки вопросов теста (вопросы с вариантами ответов).
Вариант 1.

1. Приведите основные характеристики больших данных:
а) Virtualization, Volume, Variability, Vehicle;
б) Variety, Velocity, Volume, Value;
в) Verification, Volume, Velocity, Visualization;
г) Video, Value, Variety, Volume.

2. Расставьте в правильном порядке основные этапы процесса Data Science:
а) назначение цели исследования, сбор данных, подготовка данных, исследование данных, моделирование данных, отображение данных;
б) назначение цели исследования, сбор данных, подготовка данных, моделирование данных, исследование данных, отображение данных;
в) назначение цели исследования, подготовка данных, сбор данных, моделирование данных, исследование данных, отображение данных;
г) назначение цели исследования, сбор данных, подготовка данных, отображение данных, исследование данных, моделирование данных.

3. Поясните понятие:
Hadoop представляет собой…
а) набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах;
б) распределённую СУБД, позволяющую обрабатывать большие данные;
в) язык выполнения заданий в парадигме MapReduce;
г) распределённую файловую систему для организации хранения файлов большого объёма.

4. Принцип MapReduce состоит в том, чтобы
а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-па**р**аллельной обработки.

Технология проведения: тест состоит из 50 вопросов. Вариант теста выбирается, исходя из номера зачетки (последней цифры). Время тестирования составляет 45 минут. Результаты теста проверяются по ключу правильных ответов.

Критерии оценивания:
- оценка «зачетно» выставляется студенту, если студент дал правильные ответы на 75 и более процентов заданий (тест пройден);
- оценка «не зачтено» - даны правильные ответы на менее чем 75 процентов заданий (тест не пройден).

Для оценивания результатов обучения на экзамене используется 4-балльная шкала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно».

Соотношение показателей, критериев и шкалы оценивания результатов обучения:

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Итоговый тест зачтен. Посещение лекций базового блока составляет 90%, посещение блока дополнительных лекций 50% и более. Выполнены лабораторные работы №1-7 и контрольная работа зачтена.	Повышенный уровень	Отлично
Итоговый тест зачтен. Посещение лекций базового блока составляет 90%, посещение блока дополнительных лекций составляет менее 50%. Выполнены лабораторные работы №1-5 и одна работа из №6-7, контрольная работа зачтена.	Базовый уровень	Хорошо
Итоговый тест зачтен. Посещение лекций базового блока составляет 90%. Выполнены лабораторные работы №1-5.	Пороговый уровень	Удовлетвори-тельно
Итоговый тест не зачтен и/или посещение лекций базового блока составляет менее 90% и/или не выполнены лабораторные работы №1-5.	–	Неудовлетвори-тельно

20.3 Фонд оценочных средств сформированности компетенций студентов, рекомендуемый для проведения диагностических работ

Вопросы с вариантами ответов (закрытые)

ПК-4 Способен разрабатывать комплекс требований к программному обеспечению, осуществлять его проектирование с учетом особенностей предметной области для решения прикладных задач в естественных науках, промышленности и бизнесе и управлять работами по созданию (модификации) и сопровождению информационных ресурсов

1. В чем заключается научное и общественное значение больших данных:
а) возможность извлечь экономическую выгоду;
б) новые знания о мире;
в) возможность ответить на давно интересующие вопросы;
г) возможность управления будущим.
Ответ: а.

2. Основные механизмы реализации баз данных типа NoSQL:
а) репликация и шардинг;
б) шардинг и поддержка map/reduce;
в) репликация и поддержка map/reduce;
г) репликация и горизонтальное масштабирование.
Ответ: а.

3. Какие модули входят в основной пакет для работы с большими данными ПО Hadoop:
а) Common, HDFS, MapReduce, YARN;
б) MapReduce, HDFS, YARN;
в) MapReduce, HDFS, Common, HBase;
г) Common, YARN, HBase.
Ответ: а.

4. В чем заключаются основные идеи баз данных NoSQL:

а) возможность применения к неструктурированным и слабоструктурированным данным, отсутствие SQL-запросов;
б) возможность работы с различными типами хранилищ, реаляционная модель данных, удобство для разработчиков;
в) нереляционная модель данных, закрытый исходный код, вертикальная масштабируемость;
г) нереляционная модель данных, открытый исходный код, хорошая горизонтальная масштабируемость.
Ответ: г.

5. Выберите наиболее подходящее определение термина «Hadoop»:
а) платформа для запуска приложений аналитической обработки данных;
б) ПО, предназначенное для создания и запуска распределённых приложений, обрабатывающих большие объёмы данных;
в) ПО, предназначенное для организации работы с большими данными;
г) СУБД для неструктурированных и слабоструктурированных данных.
Ответ: б.

6. Перечислите системные требования к ПО Hadoop:
а) ОС Linux, поддержка Java API, кластерная топология;
б) ОС Linux, C++;
в) ОС Windows 64-разрядная, поддержка реляционных БД;
г) любая ОС, ограничений на языки программирования нет.
Ответ: а.

7. Основные режимы запуска ПО Hadoop:
а) автономный, псевдораспределенный, полностью распределенный;
б) псевдораспределенный, распределенный;
в) автономный и псевдораспределенный;
г) локальный и сетевой.
Ответ: а.

8. Условия реализуемости концепции работы с данными MapReduce:
а) наличие распределенной файловой системы, планировщика, неиндексированное хранение данных, автоматизации распараллеливания задач на кластере;
б) поддержка репликации данных, индексированное хранение данных, наличие планировщика;
в) наличие планировщика, индексированное хранение данных, использование распределенной файловой системы;
г) наличие распределенной файловой системы, планировщика, неиндексированное хранение данных.
Ответ: а.

9. Принцип MapReduce применительно к технологии обработки больших данных состоит в том, чтобы:
а) производить вычисления на узлах, где информация изначально была сохранена;
б) использовать вычислительные мощности систем хранения;
в) использовать функциональное программирование для решения задач массивно-параллельной обработки
г) разделять узлы на те, где хранятся данные и те, на которых производятся вычисления.
Ответ: а.

10. Какая концепция положена в основу распределенной файловой системы HDFS:
а) производить вычисления на узлах, где информация изначально была сохранена;
б) однократной записи и многократного чтения;
в) возможности репликации данных;
г) распараллеливания процессов обработки данных.
Ответ: б.

11. Архитектура модуля YARN ориентирована на работу:
а) только MapReduce-программ;
б) любых распределённых приложений;
в) MapReduce-программ, разработанных для потоковых данных;
г) как MapReduce-программ, так и любых других распределённых приложений, поддерживающие соответствующие программные интерфейсы.
Ответ: г.

ПК-6 Способен проводить обработку и анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры

12. Укажите основные источники больших данных:
а) мобильные устройства;
б) машинные данные и интернет;
в) интернет и социальные сети;
г) машинные данные и мобильные устройства.
Ответ: г.

13. Позволяют ли технологии big data работать с неструктурированными или плохо структурированными данными?
а) да;
б) нет.
Ответ: а.

14. Какие характеристики определяют принцип «Трёх V» в отношении больших данных:
а) Virtualization, Volume, Variability;
б) Variety, Volume, Value;
в) Verification, Velocity, Visualization;
г) Volume, Variety, Velocity.
Ответ: г.

15. Укажите основные виды данных, которые рассматриваются в контексте работы с большими данными:
а) структурированные, полуструктированные, квазиструктурированные и неструктурированные;
б) структурированные, полуструктурированные и неструктурированные;
в) структурированные и неструктурированные;
г) неструктурированные, квазиструктурированные, неоструктрированные и структурированные.
Ответ: а.

16. Отметьте высказывания, наиболее полно характеризующие структурированные данные:

а) не зависят от модели данных, удобны для анализа, примером является аудио файлы;
б) зависят от модели данных, имеют определенную структуру, произвольный фрагмент текста трудно подвергается расшифровке;
в) данные содержат специальные теги и иные маркёры, позволяющие отделить семантические элементы, удобны для анализа;
г) зависят от модели данных, упорядочены специальным образом, обычно такие данные хранятся в виде таблиц в реляционных базах данных, удобны для анализа.
Ответ: г.

17. Укажите типы корпоративных данных:
а) машинные данные, естественные данные, социальных сетей сотрудников и клиентов;
б) фактографические, нормативно-справочные и внутренние;
в) конфиденциальные, из открытых и условно-открытых источников;
г) открытые и закрытые.
Ответ: б.

18. Выберите наиболее верное утверждение «Информация, содержащая конкретные фактические сведения о конкретных фактах, о фактических событиях, характеризующие некоторый объект и позволяющие провести сопоставление его с аналогами - это»:
а) фактографическая информация;
б) справочная информация;
в) нормативно-справочная информация;
г) документально-подтвержденная.
Ответ: а.

19. При использовании технологий big data какие данные характеризуют как «данные высокого качества»:
а) данные, не нуждающиеся в очистке;
б) данные, загруженные из достоверного источника;
в) данные, содержащие некритичные ошибки, которые не мешают их загрузке в хранилище данных;
г) все вышеперечисленные.
Ответ: а.

20. Что является критерием оценки качества данных при использовании технологий big data:
а) критичность ошибок
б) степень детализации данных
в) достоверность источника
г) скорость предоставления данных
Ответ: а.

21. Верно ли утверждение, что «При визуализации источников данных основную трудность вызывает отображение данных из реляционных баз данных, т.к. метаданные и данные отделены друг от друга»?
а) верно;
б) неверно.
Ответ: а.

22. Обязательно ли при использовании технологий big data приводить данные к единому формату?
а) да;
б) нет.
Ответ: б.

23. При применении технологий big data метаданные:
а) должны размещаться вместе с данными;
б) должны размещаться отдельно от данных;
в) не используются.
Ответ: а.

24. Укажите основные фазы моделирования данных, которые имеют место быть в работе с большими данными:
а) выбор переменных, определение значимости переменных, выполнение модели, исследование результатов;
б) выбор модели, выполнение модели, оценка результатов;
в) выбор модели и переменных, выполнение модели, определение степени соответствия модели, диагностика модели;
г) выбор модели и переменных, выполнение модели, диагностика и сравнение моделей.
Ответ: г.

25. Укажите уровни интеграции данных:
а) семантический и синтаксический;
б) физический и логический;
в) внешний и внутренний;
г) ручной и машинный.
Ответ: б.

26. Как называются технологии, позволяющие работать с большими объемами неструктурированных и плохо структурированных данных разного формата?
а) big data;
б) SQL;
в) «последующего поколения»;
г) NGN.
твет: а.

27. Выберите группы визуализаторов, используемых при работе с большими данными:
а) общего назначения, для оценки качества моделей, для интерпретации результатов анализа;
б) для оценки качества входных данных, оценки качества моделей, оценки прогнозируемых значений;
в) общего назначения, специализированные;
г) для оценки входных данных, для оценки качества моделей, для интерпретации результатов анализа.
Ответ: а.

28. Что включает в себя этап сбора данных в процессе изучения данных DS (Data Science):

а) определение источников данных, определение методов сбора данных, сбор данных, первичный анализ данных;
б) определение источников данных, формирование цепочек жизненного цикла данных и определение методов сбора данных, сбор данных, первичный анализ данных;
в) определение источников данных, формирование цепочек жизненного цикла данных и определение методов сбора данных, первичный анализ данных;
г) определение источников данных, определение методов сбора данных, сбор данных.
Ответ: б.

29. Основная задача этапа подготовки данных при реализации процесса изучения данных DS (Data Science):
а) проведение предварительного исследования данных, описание данных;
б) очистка данных и составление их описания;
в) объединение данных из разных источников и приведение их к единому формату:
г) данные разных наборов приводят к общему формату, убирают опечатки и различные ошибки ввода.
Ответ: г.

30. Дайте определение процессу преобразования данных при изучении данных DS (Data Science):
а) процесс приведения данных к виду, подходящему для моделирования данных;
б) процесс очистки данных с сокращением объема файла;
в) процесс выборки из данных полезной информации;
г) процесс приведения данных к формату, пригодному для применения SQL-запросов.
Ответ: а.

31. Расставьте в правильном порядке основные этапы процесса Data Science:
а) назначение цели исследования, сбор данных, подготовка данных, моделирование данных, исследование данных, отображение данных;
б) назначение цели исследования, сбор данных, подготовка данных, исследование данных, моделирование данных, отображение данных;
в) назначение цели исследования, подготовка данных, сбор данных, моделирование данных, исследование данных, отображение данных;
г) назначение цели исследования, сбор данных, подготовка данных, отображение данных, исследование данных, моделирование данных.
Ответ: б.

32. Основные этапы методологии аналитики данных CRISP-DM (в порядке исполнения):
а) бизнес-анализ, анализ данных, подготовка данных, моделирование, оценка решений, внедрение;
б) отбор данных, исследование отношений в данных, модификация данных, моделирование взаимозависимостей;
в) формирование бизнес-задачи, анализ-данных, сбор и подготовка данных, оценка результатов;
г) анализ источников данных, сбор данных, построение моделей, оценка моделей, внедрение.
Ответ: а.

33. На каком этапе методологии аналитики данных CRISP-DM применяется метод A/B-тестирования:
а) моделирование;
б) оценка полученных моделей;
в) оценка решений;
г) внедрение.
Ответ: а.

34. Дайте определение процесса ETL, используемого при обработке больших массивов данных:
а) комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных;
б) ПО для извлечения данных из реляционных БД и преобразование их к виду, удобному для хранения в БД NoSQL;
в) комплекс методов, очистки данных, извлекаемых из различных источников;
г) ПО для приведения данных из разных источников к единому формату.
Ответ: а.

35. Какие виды атрибутов используются для обозначения данных в выборках после проведения очистки согласно методологии аналитики данных CRISP-DM:
а) качественные атрибуты, исправленные атрибуты и забракованные;
б) качественные и некачественные атрибуты;
в) высокое качество, низкое качество;
г) допустимое качество, исправлению не подлежит.
Ответ: а.

36. Верно ли утверждение «Основная задача визуализации данных – это представление результатов ключевым участникам проекта и построение приложений на их основе»?
а) верно;
б) неверно.
Ответ: а.

37. Отметьте основные типы инструментальных средств платформ работы с данными BI (Business Intelligence):
а) средства сбора и представления информации и средства интеграции;
б) средства сбора информации, средства очистки информации, средства анализа;
в) средства сбора информации, средства преобразования к виду, удобному для обработки, средства моделирования;
г) средства представления информации, средства интеграции, средства анализа.
Ответ: г.

38. Разделение базы данных на отдельные части называется:
а) партицирование;
б) пакетирование;
в) шардинг;
г) масштабирование.
Ответ: а.

39. Что является результатом аналитического исследования?
а) новые знания;

б) новые данные;
в) новые метаданные;
г) новые переменные.
Ответ: а.

40. Для работы с большими данными в организации должна быть ИТ-инфраструктура, поддерживающая распределенное хранение данных?
а) да;
б) нет.
Ответ: а.

41. Интеллектуальный анализ данных KDD является:
а) итеративным процессом;
б) неитеративным процессом;
в) интерактивным;
г) каскадным процессом.
Ответ: а.

Вопросы с кратким текстовым ответом (открытые)

ПК-4 Способен разрабатывать комплекс требований к программному обеспечению, осуществлять его проектирование с учетом особенностей предметной области для решения прикладных задач в естественных науках, промышленности и бизнесе и управлять работами по созданию (модификации) и сопровождению информационных ресурсов

42. Как называются базы данных, в которых используются не только SQL-запросы? Ответ запишите латинскими буквами в верхнем регистре.
Ответ: NOSQL.

43. Приведите аббревиатуру распределенной файловой системы хранения данных для ПО Hadoop. Ответ запишите латинскими буквами в верхнем регистре.
Ответ: HDFS.

44. Основные режимы запуска ПО Hadoop для работы с большими данными (выберите нужные варианты и запишите ответ в виде последовательности цифр без пробела, например «35»):

1. Автономный
2. Псевдораспределенный
3. Полностью распределенный
4. Локальный
5. Сетевой.

Ответ: 123.

ПК-6 Способен проводить обработку и анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры

45. Расставьте в нужном порядке шаги по реализации алгоритма машинного обучения согласно этапу «Моделирования» методологии CRISP-DM. Запишите ответ в виде последовательности цифр без пробела, например «3412».

1) Выбрать методику моделирования

2) Сделать тесты для модели
3) Построить модель
4) Оценить модель

Ответ: 1234.

ПК-6.3 Проводит аналитическое исследование в соответствии с согласованными требованиями

46. Как называется техника масштабирования при работе с данными, которая заключается в разделении базы данных на отдельные части так, чтобы каждую из них можно было вынести на отдельный сервер? Ответ запишите русскими буквами в нижнем регистре.
Ответ: шардинг.

47. Как обозначается комплекс методов, реализующих процесс переноса исходных данных из различных источников в аналитическое приложение или поддерживающее его хранилище данных? Ответ запишите латинскими буквами в верхнем регистре.
Ответ: ETL.

48. Какие процедуры можно отнести к генерации данных при подготовке датасетов для машинного обучения или при аналитическом исследовании? Выберите нужные варианты и запишите ответ в виде последовательности цифр без пробела, например «35».
1. Агрегацию атрибутов (расчет sum, avg, min, max, var и т.д.)
2. Нормализацию атрибутов (feature scaling)
3. Заполнение пропущенных данных (missing data imputation)
4. Копирование данных и наращивание полученными наборами выборки
Ответ: 123.

49. Как называется методология по исследованию данных, в которой жизненный цикл данных включает этап моделирования данных? Ответ запишите латинскими буквами в верхнем регистре.
Ответ: CRISP (допускается CRISP-DM).

50. Отчет о результатах аналитического исследования может быть представлен в виде… Выберите варианты из списка и запишите ответ в виде последовательности цифр без пробела, например «35».
1. Презентации
2. Аналитической записки
3. Статьи
4. Доклада
Ответ: 1234.

51. Интеллектуальный анализ данных, также широко известный как обнаружение знаний из данных обозначается… Ответ запишите латинскими буквами в верхнем регистре.
Ответ: KDD.

Критерии и шкалы оценивания заданий ФОС:

Для оценивания выполнения заданий используется балльная шкала:

1) закрытые задания (тестовые с вариантами ответов, средний уровень сложности):

- 1 балл – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

2) открытые задания (тестовые с кратким текстовым ответом, повышенный уровень сложности):

- 2 балла – указан верный ответ;
- 0 баллов – указан неверный ответ (полностью или частично неверный).

Задания раздела 20.3 рекомендуются к использованию при проведении диагностических работ с целью оценки остаточных результатов освоения данной дисциплины (знаний, умений, навыков).