

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.

21.06.2021 г



**РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.В.03 Анализ больших данных**

1. Код и наименование направления подготовки / специальности:
02.04.02 Фундаментальная информатика и информационные технологии
2. Профиль подготовки / специализация/магистерская программа:
Машинное обучение и интеллектуальные информационные технологии
3. Квалификация (степень) выпускника: магистр
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 № 10 от 15.06.2021
8. Учебный год: 2022/2023 Семестр(ы): 3

9. Цели и задачи учебной дисциплины

Цели дисциплины: освоение основных технологий, методов и алгоритмов работы с большими массивами данных, которые позволяют обрабатывать, интерпретировать, оформлять и представлять профессиональному обществу результаты исследований.

Задачами курса являются:

- знакомство с основными процессами консолидации, анализа, обработки и управления больших данных;
- изучение (исследование и испытание) методов, моделей, алгоритмов и инструментальных средств работы с большими данными;
- освоение основных инструментов для аналитической работы с большими данными.

10. Место учебной дисциплины в структуре ОПОП: (цикл, к которому относится дисциплина, требования к входным знаниям, умениям и навыкам, дисциплины, для которых данная дисциплина является предшествующей)

Дисциплина относится к обязательным дисциплинам части, формируемой участниками образовательных отношений части Блока 1. блока Б1 учебного плана. Для изучения курса необходимы знания в области ИТ-технологий, в частности, по обработке, хранению и визуализации данных.

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-3	Способен обрабатывать, интерпретировать, оформлять и представлять профессиональному обществу результаты проведенных исследований.	ПК-3.1	Проводит информационный поиск для решения исследовательских задач с использованием открытых источников информации и специализированных баз данных	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.
		ПК-3.2	Анализирует и обрабатывает информацию по тематике исследований.	
ПК-5	Способен совершенствовать и разрабатывать новые методы, модели, алгоритмы, технологии работы с большими данными.	ПК-5.2	Проводит исследования и испытания разработанных методов, моделей, алгоритмов, технологий и инструментальных средств.	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
		ПК-5.3	Проводит аналитические работы на основе технологий больших данных и машинного обучения.	

				инструментальные средства для работы с данными, в том числе, с большими данными; проводить исследования и испытания методов, моделей, алгоритмов и инструментальных средств работы с большими данными. Владеть навыками инсталляции и настройки ПО для работы с большими данными.
--	--	--	--	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

12. Объем дисциплины в зачетных единицах/часах в соответствии с учебным планом —5/180

Форма промежуточной аттестации экзамен.

13. Трудоемкость по видам учебной работы

Вид учебной работы	Трудоемкость (часы)			
	Всего	В том числе в интерактивной форме	По семестрам	
			№ сем. 3	№ сем.
Аудиторные занятия			
в том числе: лекции	32		32	
практические	-		-	
лабораторные	16		16	
Самостоятельная работа	96		96	
Форма промежуточной аттестации	Экз.		Экзамен	
Итого:	180		180	

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.	Термины и определения. Особенности технологий. Сферы применения, состояние и перспективы развития.	Анализ больших данных (02.04.02, Ухлова В.В.)
1.2	Технологии консолидации, обработки и управления большими данными	Платформа Hadoop: архитектура и принцип работы. Организация файловой системы HDFS. Программный интерфейс Map Reduce. Система YARN.	
1.3	Основные процессы в Data science	Процессы сбора, подготовки и исследования данных. Методы моделирования данных. Визуализация данных.	
2. Лабораторные работы			
2.1	Методы работы с данными	Определение целей исследования, формирование ТЗ, выбор методов реализации. Сбор данных. Проверка	Анализ больших

		качества данных. Очистка данных. Выбор средств хранения данных. Методы обработки и анализа данных. Инструменты управления данными.	данных (02.04.02, Ухлоva В.В.)
--	--	---------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела) дисциплины	Виды занятий (часов)				Всего
		Лекции	Практические	Лабораторные	Самостоятельная работа	
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	24	44
4	Методы работы с данными	2	-	12	56	70
	Итого:	32	-	16	96	180

14. Методические указания для обучающихся по освоению дисциплины

Дисциплина реализуется по тематическому принципу, каждая тема представляет собой завершённый раздел курса. 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, закреплённых компетенций по ОПОП. Они организовываются в виде выполнения отдельных заданий.

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.

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. Ухлоva .— 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	Ильин, В. В. Проектный менеджмент : учебное пособие / В. В. Ильин. — 3-е изд. (эл.). — Москва : Интермедиа, 2018. — 266 с. — ISBN 978-5-91349-054-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/114754 — Режим доступа: для авториз. пользователей.

б) дополнительная литература:

№ п/п	Источник
3	Литвин, Ю. И. Проектный менеджмент: теория и практика : учебное пособие / Ю. И. Литвин, Л. И. , Р. Р. Харисова. — Москва : Прометей, 2020. — 240 с. — ISBN 978-5-907166-99-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/165992 . — Режим доступа: для авториз. пользователей.

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

№ п/п	Ресурс
4	Электронно-библиотечная система «Лань» - Режим доступа: https://e.lanbook.com
5	Электронный каталог Научной библиотеки Воронежского государственного университета. — Режим доступа: http://www.lib.vsu.ru .
6	Анализ больших данных (02.04.02, 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

Для самостоятельно подготовки обучающийся пользуется конспектами лекций и литературой по тематике лекционного материала, заданий контрольных и лабораторных работ.

Самостоятельная работа обучающегося должна включать подготовку к тестированию, лабораторным занятиям и подготовку к промежуточной аттестации. 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 на отдельном 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. (MS Office, МойОфис, LibreOffice);
- ПО Adobe Reader;
- специализированное ПО (ПО MatLab);
- интернет-браузер (Google Chrome, Mozilla Firefox).

19. Фонд оценочных средств:

№ п/п	Наименования раздела дисциплины	Компетенция(и)	Индикатор(ы) достижения компетенции	Оценочные средства
1	Понятие Data science и Big Data, область применения.	ПК-3	ПК-3.1	Тест
2	Технологии консолидации, обработки и управления большими данными.	ПК-3	ПК-3.1, ПК-3.2	Лабораторная работа 1
3	Основные процессы в Data science.	ПК-3, ПК-5	ПК-3.1, ПК-5.3	Лабораторная работа 2
4	Методы работы с данными	ПК-5	ПК-5.2	Лабораторная работа 3
Промежуточная аттестация, форма контроля - экзамен				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:

- тест,
- лабораторная работа.

Перечень заданий для лабораторных работ.

Лабораторная работа №1

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 №2

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.

Лабораторная работа №3

Пример задания.

Установить ПО Nadoor для организации работы с большими данными. При настройке ПО реализовать автономный режим работы.

Технология проведения

Лабораторная работа выполняется в учебной лаборатории. Студенту предоставляется доступ к системным настройкам ПК. Студент проводит установку системных файлов, настройку конфигурации ПО и запускает ПО в автономном режиме. Преподаватель проверяет факт установки и готовность ПК к дальнейшей работе. По окончании лабораторной работы рекомендовано восстановление системы до первоначального состояния.

Критерии оценивания:

- оценка «отлично» выставляется студенту, если все этапы установки ПО пройдены, ПО настроено и готово к работе;
- оценка «хорошо» - если все этапы установки ПО пройдены, но ПО не настроено;
- оценка «удовлетворительно» - если студент не смог пройти все этапы установки ПО;
- оценка «неудовлетворительно» - работа не выполнена.

Тестовые задания.

Пример компоновки вопросов теста (вопросы с вариантами ответов).

Вариант 1.

1. Приведите основные характеристики больших данных:

- a) Virtualization, Volume, Variability, Velocity;
- 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:

- a) 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...

- a) 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

- a) 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: 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 «отлично» выставляется студенту, если студент дал правильные ответы на 90 и более процентов заданий (тест пройден);
- оценка «хорошо» выставляется студенту, если студент дал правильные ответы менее, чем на 90 и более 80 процентов заданий (тест пройден);
- оценка «удовлетворительно» выставляется студенту, если студент дал правильные ответы менее 80 и более 50 процентов заданий (тест пройден);
- оценка «неудовлетворительно» - даны правильные ответы на менее чем на 50 процентов заданий (тест не пройден).

20.2 Промежуточная аттестация

Промежуточная аттестация по дисциплине осуществляется с помощью следующих оценочных средств: вопросы к зачету.

Перечень вопросов к экзамену (КИМ №1):

1. Основные типы больших данных.
2. Структура процесса Data Science.
3. Основные компоненты инфраструктуры работы с большими данными.
4. Кластер Hadoop..
5. Кластер Spark.
6. Распределенная файловая система HDFS.
7. Структуры данных, применяемые для работы с большими массивами данных
8. База данных NoSQL.
9. Сравнение баз данных SQL и NoSQL.
10. Теорема CAP.
11. Парадигма MapReduce..
12. Методы исследования больших данных.
13. Основные процессы в исследовании данных.
14. Методология ведения проектов CRISP.
15. Методология ведения проектов KDD.
16. Методология ведения проектов SEMMA.
17. Инструменты визуализации данных.
18. Алгоритм развертывания локального кластера Hadoop на платформе с ОС Windows.

Для оценивания результатов обучения на экзамене используются следующие показатели (ЗУН):

- 1) знать основные технологии консолидации, обработки и управления большими данными, позволяющие осуществлять поиск, сбор и хранение информации из открытых источников и специализированных баз данных;
- 2) знать основные методологии анализа данных; алгоритмы обработки данных.
- 3) знать основные методики исследования и испытания разработанных методов, моделей, алгоритмов, технологий и инструментальных средств по работе с данными;
- 4) уметь осуществлять информационный поиск с использованием открытых источников информации и специализированных баз данных;

- 5) уметь использовать инструментальные средства для работы с данными, в том числе, с большими данными;
- 6) уметь проводить исследования и испытания методов, моделей, алгоритмов и инструментальных средств работы с большими данными;
- 7) владеть навыками инсталляции и настройки ПО для работы с большими данными.

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
Обучающийся демонстрирует знание теоретических основ (показатели 1-3), умение применять теорию на практике (показатели 4-6), По лабораторным работам получены оценки «отлично» и «хорошо». По тесту набрано более 90% правильных ответов.	Повышенный уровень	Отлично
Ответ на контрольно-измерительный материал не соответствует одному (двум) из показателей ЗУН, но обучающийся дает правильные ответы на дополнительные вопросы. По лабораторным работам получены оценки «отлично» и «хорошо». По тесту набрано более 80% правильных ответов.	Базовый уровень	Хорошо
Ответ на контрольно-измерительный материал не соответствует любым трем (четырем) из показателей ЗУН, обучающийся дает неполные ответы на дополнительные вопросы, допускает ошибки в терминологии. По лабораторным работам получены оценки «отлично» или «хорошо», или «удовлетворительно». По тесту набрано более 50% правильных ответов.	Пороговый уровень	Удовлетворительно
Ответ на контрольно-измерительный материал не соответствует ни одному из показателей ЗУН. Обучающийся не владеет терминологией данной области знаний. Задание лабораторных работ не выполнено.	–	Неудовлетворительно