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

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ
Б1.В.01 Основы Big Data

1. Код и наименование направления подготовки / специальности:

38.04.05 Бизнес-информатика

2. Профиль подготовки / специализация/магистерская программа:

Информационная бизнес-аналитика

3. Квалификация (степень) выпускника: магистр

4. Форма обучения: заочная

5. Кафедра, отвечающая за реализацию дисциплины: математических методов исследования операций

6. Составители программы: Ухлова В.В., к.ф.-м.н, доцент кафедры математических методов исследования операций

7. Рекомендована: НМС факультета Прикладной математики, информатики и механики № 10 от 15.06.2021

8. Учебный год: 2022/2023 **Сессия:** 5, 6

9. Цели и задачи учебной дисциплины

Цели дисциплины: освоение основных технологий, методов и алгоритмов работы с большими массивами данных, которые позволяют обрабатывать, анализировать, интерпретировать, оформлять и представлять профессиональному обществу результаты исследований, позволяют разрабатывать профессионально--ориентированные информационные системы с учетом возможностей современных интеллектуальных информационных технологий.

Задачами курса являются:

- знакомство с основными процессами консолидации, анализа, обработки и управления больших данных;
- изучение и совершенствование методов, алгоритмов и инструментальных средств работы с большими данными для возможности проведения аналитических исследований в рамках профессиональной деятельности;
- освоение основных навыков ведения проектов в области больших данных, в том числе, по созданию и внедрению профессионально-ориентированных информационных систем с учетом возможностей современных интеллектуальных информационных технологий.

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

Дисциплина относится к обязательным дисциплинам вариативной части программы обучения. 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 проводить обработку и анализ больших данных на базе современных языков программирования и пакетов прикладных программ моделирования	ПК-3.1	Организует сбор данных и проводит аналитическое исследование в соответствии с согласованными требованиями	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
		ПК-3.2	Разрабатывает и совершенствует методы анализа массовых количественных и нечисловых данных на базе современных языков программирования и технологий управления данными	
ПК-4	Способен управлять разработкой профессионально--ориентированных информационных	ПК-4.3	Организует работы по созданию и внедрению профессионально-ориентированных информационных систем с учетом возможностей	

	систем с учетом возможностей современных интеллектуальных информационных технологий		современных интеллектуальных информационных технологий	специализированных баз данных; использовать 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 —4/144

Форма промежуточной аттестации *экзамен*.

13. Трудоемкость по видам учебной работы

Вид учебной работы	Трудоемкость (часы)				
	Всего	В том числе в интерактивной форме	По семестрам/ сессиям		
			№ сессии. 5	№ сессии 6	…..
Аудиторные занятия					
в том числе: лекции	8	8	-	8	
практические		-			
лабораторные	8	6	2	8	
Самостоятельная работа	119	54	65	119	
Форма промежуточной аттестации	9	0/0	0/9	0/9	
Итого:	144	68	76	144	

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.	Основы технологий Big Data (38.04.05.)
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 качества данных. Очистка данных. Выбор средств хранения данных. Методы обработки и анализа данных. Инструменты управления данными.	Основы технологий Big Data(38.04.05.)

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	2	-	-	6	8
2	Основные процессы в Data science	2	-	-	20	22
3	Технологии консолидации, обработки и управления большими данными	4	-	-	42	37
4	Методы работы с данными		-	8	60	78
Контроль						9
	Итого:	8	-	8	128	144

14. Методические указания для обучающихся по освоению дисциплины

Дисциплина реализуется по тематическому принципу, каждая тема представляет собой завершенный 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, закрепленных компетенциями 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.

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

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.

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	Основы технологий Big Data (38.04.05, 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 в себя 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 b 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-3.2	Контрольная работа, 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-3.1	Лабораторная работа 2
4	Методы работы с данными	ПК-4	ПК-4.3	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:

- контрольная работа,
- тест,
- 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.

Примерный перечень заданий контрольной работы

1. Приведите основные характеристики больших данных.
2. Расставьте в правильном порядке основные этапы процесса Data Science: назначение цели исследования, подготовка данных, сбор данных, моделирование данных, исследование данных, отображение данных.
3. Дайте определение «Hadoop».
4. Поясните принцип MapReduce
5. Поясните, в чем заключается научное и общественное значение больших данных.
6. Что включает в себя экосистема Big Data.
7. Укажите основные навыки и умения специалиста Data Science
8. Прокомментируйте основные режимы запуска ПО Hadoop.
9. Какие виды узлов включает в себя ядро кластера Hadoop.
10. Укажите назначение Job Tracker в реализации кластера Hadoop.
11. Какая концепция положена в основу HDFS.
12. Поясните основной принцип хранения данных в HDFS.
13. Приведите условия реализуемости концепции MapReduce.
14. Дайте определение БД NoSQL.

Технология проведения

Контрольная включает в себя 30 вопросов, вариант вбирается исходя из номера зачетки (последней цифры). Время выполнения рассчитывается из соотношения 10 вопросов – 30 минут.

Критерии оценивания:

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

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-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 «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).

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 Hadoop 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 инсталляцию 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 инсталляции ПО пройдены, ПО настроено и готово к работе;
- оценка «хорошо» - если все этапы инсталляции ПО пройдены, но ПО не настроено;
- оценка «удовлетворительно» - если студент не смог пройти все этапы инсталляции ПО;
- оценка «неудовлетворительно» - работа не выполнена.

20.2 Промежуточная аттестация

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

Перечень вопросов к экзамену:

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:

1) знание основных технологий консолидации, обработки и управления большими данными, позволяющих осуществлять поиск, сбор и хранение информации из открытых источников и специализированных баз данных;
2) знание основных методологий анализа данных; алгоритмов обработки данных;

3) знание основных методик исследования и испытаний разработанных методов, моделей, алгоритмов, технологий и инструментальных средств по работе с данными;
4) умения осуществлять информационный поиск с использованием открытых источников информации и специализированных баз данных;
5) умение использовать инструментальные средства для работы с данными, в том числе, с большими данными;
6) умения проводить исследования и испытания методов, моделей, алгоритмов и инструментальных средств работы с большими данными;
7) владение 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 или контрольной работе 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 или контрольной работе 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 или контрольной работе набрано более 50% правильных ответов.	Пороговый уровень	Удовлетвори-тельно
Ответ на контрольно-измерительный материал не соответствует ни одному из показателей ЗУН. Обучающийся не владеет терминологией данной области знаний. Задание лабораторных работ и/или тест (контрольная работа) не выполнены.	–	Неудовлетвори-тельно