Минобрнауки России

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ВГУ»)

УТВЕРЖДАЮ

Заведующий кафедрой

Сирота Александр Анатольевич

Кафедра технологий обработки и защиты информации

23.04.2024

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

Б1.В.05 Анализ больших данных

1. Код и наименование направления подготовки/специальности:

09.04.02 Информационные системы и технологии

2. Профиль подготовки/специализация:

Системы прикладного искусственного интеллекта

3. Квалификация (степень) выпускника:

Магистратура

4. Форма обучения:

Очная

5. Кафедра, отвечающая за реализацию дисциплины:

Кафедра технологий обработки и защиты информации

6. Составители программы:

Гаршина Вероника Викторовна, к.т.н., доцент

7. Рекомендована:

№5 от 05.03.2024

8. Учебный год:

2025-2026

9.Цели и задачи учебной дисциплины:

Целями освоения дисциплины является формирование у студентов профессиональных компетенций в области разработки и использования систем обработки и анализа больших массивов данных (Big Data).

Основные задачи дисциплины:

- ознакомление с базовыми понятиями Big Data – аналитики, изучение основных моделей представления больших данных, современных методов и технологий, применяемых в системах анализа больших данных;
- изучение и практическое освоение методологий и методик анализа и прогнозирования в Big Data;
- освоение студентами современных технологий, применяемых в области создания и обслуживания больших данных.

10. Место учебной дисциплины в структуре ООП:

Дисциплина относится к дисциплинам вариативной части базового модуля учебного плана Б1.В.

Для успешного освоения необходимо предварительное изучение следующих дисциплин: математические методы в современных информационных технологиях, нейронные сети и глубокое обучение, системы поддержки принятия решений, нечеткие модели и алгоритмы принятия решений.

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

Код и название компетенции	Код и название индикатора компетенции	Знания, умения, навыки
ПК-4 Способен проектировать архитектуру программного средства	ПК-4.1 Умеет определять состав компонентов программного средства	**Знает** терминологию, базовые понятия и технологии Big Data,особенности работы с большими данными разных типов и природы (структурировнными, полуструктурированными, неструктурированными) . **Умеет** применять современные технологии и инструменты в области обработки больших данных к практическим задачам в различных предметных областях для понимания их внутренних связей и процессов в исследуемых системах.
ПК-6 Способен определять качество проводимых исследований, обрабатывать, интерпретировать и оформлять результаты проведенных исследований и представлять ре-зультаты профессиональному сообществу	ПК-6.1 Умеет обрабатывать данные проводимых исследований с использованием современных методов анализа информации и информационных технологий	**Умеет** проводить аналитические исследования на больших объемах данных, с использованием современных методов анализа информации, применять современные технологии и инструменты в области обработки больших данных к практическим задачам в различных предметных областях для понимания их внутренних связей и процессов в исследуемых системах.

12. Объем дисциплины в зачетных единицах/час:

3/108

Форма промежуточной аттестации:

Зачет с оценкой

13. Трудоемкость по видам учебной работы

Вид учебной работы	Семестр 3	Семестр 4	Всего
Аудиторные занятия	0	36	36
Лекционные занятия		12	12
Практические занятия			0
Лабораторные занятия		24	24
Самостоятельная работа	0	72	72
Курсовая работа			0
Промежуточная аттестация	0	0	0
Часы на контроль			0
Всего	0	108	108

13.1. Содержание дисциплины

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
1	Определение больших данных. Технологии хранения больших данных.	**Лекции по разделу** 1.Большие данные (big data) в информационных технологиях.Понятие Больших данных. Особенности сбора, хранения, обработки и анализа больших массивов данных. Три V: объём (volume), скорость (velocity), многообразие (variety). Источники больших данных. Использование больших данных в науке, бизнесе, государственном управлении. Подходы, инструменты и методы обработки структурированных и неструктурированных данных больших объёмов 2.Средства массово-параллельной обработки неопределённо структурированных данных - решениями категории NoSQL, алгоритмы MapReduce, программные каркасы и библиотеки проекта Hadoop. Базы данных NoSQL. Варианты построения распределенных баз данных, репликация, фрагментация. Согласованность. CAP-теорема. Классы NoSQL баз данных. Примеры СУБД NoSQL. Семейства столбцов. Графовые СУБД . **Лабораторные занятия по разделу** Лабораторная работа №1 Средства построения распределенных информационных систем для BigData. Обзор возможностей. Лабораторная работа №2 Изучение и конфигурирование программного комплекса Apache Hadoop. Размещения набора данных по заданной тематике.	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
2	Процесс анализа больших данных. Технологии анализа больших данных.	**Лекции по разделу** 3. Методы и техники анализа, применимые к большим данным: методы класса Data Mining: обучение ассоциативным правилам (англ. association rule learning), классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным), кластерный анализ, регрессионный анализ. Интегрирование разнородных данных из разнообразных источников для возможности глубинного анализа. Примеры цифровая обработка сигналов и обработка естественного языка (включая тональный анализ); Машинное обучение, обучение с учителем и без учителя, Ensemble learning (англ.) - использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов на основе базовых моделей (англ. constituent models, ср. со статистическим ансамблем в статистической механике); искусственные нейронные сети, сетевой анализ, оптимизация, в том числе генетические алгоритмы; распознавание образов. 4. Аналитика; имитационное моделирование; пространственный анализ (англ. Spatial analysis) - класс методов, использующих топологическую, геометрическуюи географическую информацию в данных; статистический анализ, в качестве примеров методов приводятся A/B-тестирование и анализ временных рядов; визуализация аналитических данных - представление информации в виде рисунков, диаграмм, с использованием интерактивных возможностей и анимации как для получения результатов, так и для использования в качестве исходных данных для дальнейшего анализа. **Лабораторные занятия по разделу** Лабораторная работа №3,4 Документно-ориентированные распределенные СУБД. Работа с MongoDB. Запросы на выборку и модификацию. Использование драйверов. Настройка фрагментации. Использование Map-Reduce Изучение основных классов NoSQL СУБД, графовых, мультиколоночных, документо-ориентированных, типа "имя=значение". Проектирование и разработка графовой базы данных в СУБД Neo4j на заданную тему. Поисковые запросы на языке Cipher.	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

п/п	Наименование раздела дисциплины	Содержание раздела дисциплины	Реализация раздела дисциплины с помощью онлайн-курса, ЭУМК
3	Методы прогнозирования.	5.Понятие прогноза и предвидения. Отличие прогнозирования от предвидения. Закон распределения случайной величины. Статистические оценки параметров. Доверительные области. Теория моментов. Корреляционный анализ. Использование модели множественной линейной регрессии для прогнозирования показателей. Доверительные интервалы для зависимой переменной. Сглаживание временных рядов. Динамические модели с распределенными лагами. Стационарные временные ряды. Тестирование стационарности. Коинтеграция. Анализ временных рядов. 6. Адаптивные и мультипликативные методы прогнозирования. Экспоненциальное сглаживание. Авторегрессионные модели. Модели скользящего среднего. Интегрированные процессы. Идентификация авторегрессионной модели скользящего среднего. Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза. Дисперсионный анализ влияния качественных факторов. Ранговые методы. Факторный анализ. Метод главных факторов. Многомерное шкалирование. Классическая модель многомерного шкалирования. Неметрические методы. Кластерный анализ. Дискриминантный анализ. Многомерный статистический анализ. **Лабораторные занятия по разделу** Лабораторная работа №5,6 Разработка процедур анализа данных на языке R Разработка процедур анализа данных на языке Python	Создан электронный онлайн - курс, размещены материалы к практикческим и лабораторным работам.

13.2. Темы (разделы) дисциплины и виды занятий

№ п/п	Наименование темы (раздела)	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятельная работа	Всего
1	Определение больших данных. Технологии хранения больших данных.	4	0	4	12	20

№ п/п	Наименование темы (раздела)	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятельная работа	Всего
2	Процесс анализа больших данных. Технологии анализа больших данных.	4	0	12	30	46
3	Методы прогнозирования.	4	0	8	30	42
		12	0	24	72	108

14. Методические указания для обучающихся по освоению дисциплины

1) При изучении дисциплины рекомендуется использовать следующие средства: рекомендуемую основную и дополнительную литературу; методические указания и пособия; контрольные задания для закрепления теоретического материала; электронные версии учебников и методических указаний для выполнения лабораторно - практических работ (при необходимости материалы рассылаются по электронной почте).

2) Для максимального усвоения дисциплины рекомендуется проведение письменного опроса (тестирование, решение задач) студентов по материалам лекций и лабораторных работ. Подборка вопросов для тестирования осуществляется на основе изученного теоретического материала. Такой подход позволяет повысить мотивацию студентов при конспектировании лекционного материала.

3) При проведении лабораторных занятий обеспечивается максимальная степень соответствия с материалом лекционных занятий и осуществляется экспериментальная проверка методов, алгоритмов и технологий, применяемых в интеллектуальной обработке информации, излагаемых в рамках лекций.

4) При переходе на дистанционный режим обучения для создания электронных курсов, чтения лекций он-лайн и проведения лабораторно- практических занятий используется информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете.

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

№ п/п	Источник
1	Сирота, Александр Анатольевич. Методы и алгоритмы анализа данных и их моделирование в MATLAB : [учебное пособие] / А.А. Сирота .— Санкт-Петербург : БХВ-Петербург, 2016 .— 381 с. : ил. — Библиогр.: с. 371-374 .— Предм. указ.: с. 377-381 .— ISBN 978-5-9775-3778-0.

№ п/п	Источник
2	Макшанов, А. В. Большие данные. Big Data : учебник для вузов / А. В. Макшанов, А. Е. Журавлев, Л. Н. Тындыкарь. — Санкт-Петербург : Лань, 2021. — 188 с. — ISBN 978-5-8114-6810-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/165835 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
3	Дюк, В. А. Логический анализ данных : учебное пособие / В. А. Дюк. — Санкт-Петербург : Лань, 2020. — 80 с. — ISBN 978-5-8114-4180-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/126935 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.

б) дополнительная литература:

№ п/п	Источник
1	Силен Д., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных/ Д. Силен, А.Мейсман, М. Али – М.: Питер -2018. – 336 с.
2	Лэм, Ч. Hadoop в действии / Ч. Лэм. - М. : ДМК Пресс, 2012. - 424 с
3	Маккинни, У. Python и анализ данных / У. Маккинни ; перевод с английского А. А. Слинкина. — 2-ое изд., испр. и доп. — Москва : ДМК Пресс, 2020. — 540 с. — ISBN 978-5-97060-590-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/131721 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
4	Методы и модели исследования сложных систем и обработки больших данных : монография / И. Ю. Парамонов, В. А. Смагин, Н. Е. Косых, А. Д. Хомоненко ; под редакцией В. А. Смагинаи А. Д. Хомоненко. — Санкт-Петербург : Лань, 2020. — 236 с. — ISBN 978-5-8114-4006-1. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/126938 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
5	Изучаем Spark: молниеносный анализ данных / Х. Карау, Э. Конвински, П. Венделл, М. Захария. — Москва : ДМК Пресс, 2015. — 304 с. — ISBN 978-5-97060-323-9. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/90118 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
6	Осипенков, Я. М. Google Analytics 2019. Полное руководство : руководство / Я. М. Осипенков. — Москва : ДМК Пресс, 2019. — 748 с. — ISBN 978-5-97060-788-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140575 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.

№ п/п	Источник
7	Нидхем, М. Графовые алгоритмы : руководство / М. Нидхем, Э. Холдер ; перевод с английского В. С. Яценкова. — Москва : ДМК Пресс, 2020. — 258 с. — ISBN 978-5-97060-799-2. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/140578 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
8	Бонцанини, М. Анализ социальных медиа на Python. Извлекайте и анализируйте данные из всех уголков социальной паутины на Python / М. Бонцанини ; перевод с английского А. В. Логунова. — Москва : ДМК Пресс, 2018. — 288 с. — ISBN 978-5-97060-574-5. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/108129 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
9	Гулаков, В. К. Структуры и алгоритмы обработки многомерных данных : монография / В. К. Гулаков, А. О. Трубаков, Е. О. Трубаков. — 2-е изд., стер. — Санкт-Петербург : Лань, 2021. — 356 с. — ISBN 978-5-8114-7965-8. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/169812 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.

в) информационные электронно-образовательные ресурсы:

№ п/п	Источник
1	Электронный каталог Научной библиотеки Воронежского государственного университета. – (http // www.lib.vsu.ru/).
2	Образовательный портал «Электронный университет ВГУ».– (https://edu.vsu.ru/)
3	ЭБС «Университетская библиотека online» – Контракт №3010 06/11 23 от 26.12.2023 (с 26.12.2023 по 25.12.2024)
4	ЭБС «Консультант студента» – Лицензионный договор №980КС/12-2023 / 3010-06/01-24 от 24.01.2024 с 24.01.2024 по 11. 01.2025)
5	ЭБС Лань Лицензионный договор №3010, (с 01/03/2024 по 28.02.2025) 06/02 24 от 13.02.2024 (с дополнительным соглашением №1 от 14.03.2024)
6	Электронная библиотека ВГУ, Договор №ДС-208 от 01.02.2021 с ООО «ЦКБ «БИБКОМ» и ООО «Агентство «Книга-Сервис» о создании Электронной библиотеки ВГУ, (с 01.02.2021 по 31.01.2027)
7	ЭБС BOOK.ru, Договор №3010 15/983 23 от 20.12.2023, (с 01.02.2024 по 31.01.2025).

16. Перечень учебно-методического обеспечения для самостоятельной работы

№ п/п	Источник
1	Нестеров, С. А. Основы интеллектуального анализа данных. Лабораторный практикум : учебное пособие / С. А. Нестеров. — Санкт-Петербург : Лань, 2020. — 40 с. — ISBN 978-5-8114-4509-7. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/130181 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
2	Бутаков, Н. А. Обработка больших данных с Apache Spark : учебно-методическое пособие / Н. А. Бутаков, М. В. Петров, Д. Насонов. — Санкт-Петербург : НИУ ИТМО, 2019. — 50 с. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/136573 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
3	Железнов, М. М. Методы и технологии обработки больших данных : учебно-методическое пособие / М. М. Железнов. — Москва : МИСИ – МГСУ, 2020. — 46 с. — ISBN 978-5-7264-2193-3. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/145102 (дата обращения: 01.05.2024). — Режим доступа: для авториз. пользователей.
4	Информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru)

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

Для реализации учебного процесса используются:

1. ПО Microsoft в рамках подписки "Imagine/Azure Dev Tools for Teaching", договор №3010-16/96-18 от 29 декабря 2018г.
2. ПО MATLAB Classroom ver. 7.0, 10 конкурентных бессрочных лицензий на каждый, компоненты: Matlab, Simulink, Stateflow, 1 тулбокс, N 21127/VRN3 от 30.09.2011 (за счет проекта ЕК TEMPUS/ERAMIS).
3. ПО Матлаб в рамках подписки Университетская лицензия на программный комплекс для ЭВМ - MathWorks MATLAB Campus-Wide Suite по договору 3010-16/118-21 от 27.12.2021 (до 01.2025)
4. При проведении занятий в дистанционном режиме обучения используются технические и информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете.
5. Apache Spark Hadoop (http://hadoop.apache.org) – проект с открытым исходным кодом
6. MongoDB — документоориентированная система управления базами данных с открытым исходным кодом (класс БД – NoSQL) (http://www.mongodb.com).
7. СУБД Neo4j Aura (безсерверная облачная графовая СУБД).
8. Язык программирования обработки данных – R (https://www.r-project.org/).
9. Язык программирования Python.
10. При проведении занятий в дистанционном режиме обучения используются технические и информационные ресурсы Образовательного портала "Электронный университет ВГУ (https:\\edu.vsu.ru), базирующегося на системе дистанционного обучения Moodle, развернутой в университете, а также другие доступные ресурсы сети Интернет.

18. Материально-техническое обеспечение дисциплины:

1. **394018, г. Воронеж, площадь Университетская, д. 1, корпус 1а, аудитория 290**

Учебная аудитория: специализированная мебель, компьютер преподавателя Pentium-G3420-3,2ГГц, монитор с ЖК 17", мультимедийный проектор, экран. Система для видеоконференций Logitech ConferenceCam Group и ноутбук 15.6’’ FHD Lenovo V155-15API

ПО: ОС Windows v.7, 8, 10, Набор утилит (архиваторы, файл-менеджеры), LibreOffice v.5-7, Foxit PDF Reader

специализированная мебель: доска меловая 1 шт., столы 31 шт., стулья 64 шт.; выход в Интернет, доступ к фондам учебно-методической документации и электронным изданиям.

2.Компьютерный класс (один из №1-4 корп. 1а, ауд. № 382-385). Учебная аудитория: персональные компьютеры на базе i3-2120-3,3ГГц, мониторы ЖК 19’’ (16 шт.), мультимедийный проектор, экран.

ПО: ОС Windows v.7, 8, 10, Набор утилит (архиваторы, файл-менеджеры), LibreOffice v.5-7, Foxit PDF Reader

16 шт., специализированная мебель: доска маркерная 1 шт., столы 16 шт., стулья 33 шт.; доступ к фондам учебно-методической документации и электронным изданиям, доступ к электронным библиотечным системам, выход в Интернет.

19. Оценочные средства для проведения текущей и промежуточной аттестаций

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

№ п/п	Разделы дисциплины (модули)	Код компетенции	Код индикатора	Оценочные средства для текущей аттестации
1	Разделы 1-3 Определение больших данных. Технологии хранения больших данных. Процесс анализа больших данных. Технологии анализа больших данных. Методы прогнозирования.	ПК-4 Способен проектировать архитектуру программного средства	ПК-4.1 Умеет определять состав компонентов программного средства	Устный опрос, собеседование. Контрольная работа по соответствующим разделам. Лабораторные работы 1-6

№ п/п	Разделы дисциплины (модули)	Код компетенции	Код индикатора	Оценочные средства для текущей аттестации
2	Разделы 1-3 Определение больших данных. Технологии хранения больших данных. Процесс анализа больших данных. Технологии анализа больших данных. Методы прогнозирования.	ПК-6 Способен определять качество проводимых исследований, обрабатывать, интерпретировать и оформлять результаты проведенных исследований и представлять ре-зультаты профессиональному сообществу	ПК-6.1 Умеет обрабатывать данные проводимых исследований с использованием современных методов анализа информации и информационных технологий	Устный опрос, собеседование. Контрольная работа по соответствующим разделам. Лабораторные работы 1-6
3				

Промежуточная аттестация

Форма контроля - Зачет с оценкой

Оценочные средства для промежуточной аттестации

Текущий контроль успеваемости по дисциплине осуществляется с помощью следующих оценочных средств:

Устный опрос на практических занятиях

Контрольная работа по теоретической части курса

Лабораторные работы

20 Типовые оценочные средства и методические материалы, определяющие процедуры оценивания

20.1 Текущий контроль успеваемости

№ п/п	Наименование оценочного средства	Представление оценочного средства в фонде	Критерии оценки
1	Устный опрос на лабораторных занятиях	Вопросы по темам/разделам дисциплины	Правильный ответ – зачтено, неправильный или принципиально неточный ответ - не зачтено
2	Контрольные задания по разделам дисциплины, тесты	Теоретические вопросы по темам/разделам дисциплины	Шкала оценивания соответствует приведенной в разделе 20.2

3	Лабораторная работа	Содержит 6 лабораторных заданий, предусматривающие разработку, тестирование и эксплуатацию моделей и алгоритмов анализа данных с использованием различных методов обучения.	При успешном выполнении работ в течение семестра фиксируется возможность оценивания только теоретической части дисциплины в ходе промежуточной аттестации (зачета), в противном случае проверка задания по лабораторным работам выносится на экзамен.

Пример теста и контрольных заданий

Приведённые ниже задания рекомендуется использовать при проведении диагностических работ для оценки остаточных знаний по дисциплине

Компетенция ПК-4, ПК-6

1. О соотношении аналоговой и цифровой информации:

1. Большинство данных в мире в 2019 году содержалось:

i. В цифровом виде

ii. В аналоговом виде

2. В каком веке произошёл перевес объёмов накопленных человечеством данных в сторону цифровых?

3. Объём накопленных человечеством цифровых данных на 2019 год измеряется:

i. Петабайтами

ii. Зеттабайтами

iii. Экзабайтами

iv. Йоттабайтами

4. Сколько Петабайт в Зеттабайте?

2. История больших данных

1. Укажите фактор, способствовавший появлению тренда больших данных

i. Маркетинговые кампании крупных корпораций

ii. Снижение издержек на хранение данных

iii. Появление новых технологий обработки потоковых данных

iv. Выпуск баз данных с обработкой данных в памяти

2. Какие вероятные разочарования тренда больших данных?

i. Из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных.

3. Отметьте значимые события, повлиявшие на формирование тренда больших данных:

i. Разработка Hadoop

ii. Изобретение принципа MapReduce

iii. Разработка языка Pyhton

iv. Победа Deepblue в матче с Г.Каспаровым.

3. Определение больших данных:

1. Выберите верный ответ

i. Большие данные – это обработка или хранение более 1 Тб

информации.

ii. Проблема больших данных – это такая проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.

iii. Большие данные – это огромная PR-акция крупных вендоров и не более того.

iv. Большие данные – это явление, когда цифровые данные наиболее полно представляют изучаемый объект.

2. Выберите неверный ответ:

i. Большие данные – это данные объёма свыше 1 Тб

ii. Проблема больших данных – это проблема, когда при существующих технологиях хранения и обработки сущностная обработка данных затруднена или невозможна.

iii. Большие данные – это тренд в области ИТ, подогреваемый маркетинговыми кампаниями крупных вендоров.

iv. Большие данные как правило не структурированы.

3. Отметьте те из вариантов, в которых данные структурированы:

i. Данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word.

ii. Таблица с ежедневными показаниями температуры помещения за год в файле формата csv.

iii. Текст педагогической поэмы А.С. Макаренко, представленный в формате PDF.

iv. Библиотека фильмов, представленных в формате mpeg4 на

одном жестком диске.

4. Характеристики Big Data:

1. Перечислите четыре основных характеристики Big Data:

i. Virtualization, Volume, Variability, Vehicle

ii. Variety, Velocity, Volume, Value

iii. Verification, Volume, Velocity, Visualization

iv. Video, Value, Variety, Volume

2. Выберите

2. Выберите неверное высказывание:

i. Большие объёмы данных приводят к слабой их структуризации, поэтому появляется такое разнообразие данных.

ii. Увеличившаяся производительность телекоммуникационных каналов привела к росту объёмов передаваемой информации.

iii. Удешевление систем хранения на единицу информации привело к росту рынка больших данных.

iv. Большое разнообразие источников данных

3. Отметьте неверное понимание Variety в контексте характеристик

Big Data:

i. Высокая скорость генерирования данных.

ii. Разные типы данных в колонках таблиц реляционных СУБД.

iii. Разнообразие отраслей, являющихся источниками данных.

iv. Разнообразие типов данных, включающих в себя структури-

рованные, полуструктурированные и неструктурированные.

5. Принцип MapReduce

1. Принцип MapReduce состоит в том, чтобы

i. Производить вычисления на узлах, где информация изначально была сохранена

ii. Использовать вычислительные мощности систем хранения

iii. Использовать функциональное программирование для решения задач массивно-параллельной обработки

2. Выберите одно неверное высказывание про MapReduce:

i. Интерфейс для массово-параллельной обработки данных, где вычисления производятся на узлах, где информация изначально была сохранена

ii. MapReduce – это две операции: распределения и сборки данных

iii. MapReduce был придуман разработчиками Hadoop

iv. MapReduce был анонсирован разработчиками Google

3. Каков теоретический прирост производительности при подсчёте

числа слов в тексте при работе MapReduce при переходе от одного узла к двум?

6. Технологии хранения

1. Какие из следующих технологий СУБД не используют принцип MapReduce

i. Hadoop

ii. Cassandra

iii. HDInsight

iv. Redis

2. Какие СУБД полностью полагаются на оперативную память при хранении информации:

i. Oracle Exalytics

ii. SAP HANA

iii. BigTable

iv. HBase

3. В чём преимущество колоночно-ориентированных СУБД?

i. Они позволяют выполнять более сложные SQL-запросы по сравнению с реляционными СУБД

ii. Они позволяют динамически дополнять содержание записей новыми полями

iii. Они имеют более гибкие возможности аналитики.

iv. Они позволяют эффективно делать межколоночные сравнения.

7. «Песочница» в аналитическом процессе

1. Для чего аналитику необходима «песочница»?

i. Для высокопроизводительной аналитики за счёт использования оперативной памяти и inDB операций.

ii. Для хранения всех полученных от заказчика данных.

iii. Для построения отчётов о результатах анализа

iv. Для снижения затрат, связанных с репликацией данных

2. Какие из следующих средств разумно использовать для анализа данных, представленных единственным csv-файлом размера более100Гб:

i. Hadoop

ii. Data Warehouse

iii. «Песочница»

iv. Python

3. Выберите верное утверждение:

i. Data Warehouse создаются для проверки гипотез при анализе больших данных.

ii. «Песочница» используется для снижения нагрузки на основной Data Warehouse.

iii. Каждый Data Warehouse должен содержать «песочницу».

iv. «Песочница» необходима для любого процесса аналитики.

8. CRISP-DM

1. Расставьте последовательность этапов проекта аналитики в соответствии с CRISP-DM.

i. Понимание бизнеса (Business understanding)

ii. Понимание данных (Data Understanding)

iii. Подготовка данных (Data Preparation)

iv. Моделирование (Modeling)

v. Оценка (Evaluation)

vi. Внедрение (Deployment)

2. На каком из этапов процесса CRISP-DM происходит проверка гипотез?

i. Понимание бизнеса (Business understanding)

ii. Понимание данных (Data Understanding)

iii. Моделирование (Modeling)

iv. Оценка (Evaluation)

3. Вы являетесь владельцем и аналитиком в компании из 10 человек, в которой требуется проанализировать продажи за 1 год (1 млн. продаж). Какие из этапов CRISP-DM можно опустить:

i. Понимание бизнеса (Business understanding)

ii. Подготовка данных (Data Preparation)

iii. Моделирование (Modeling)

iv. Оценка (Evaluation)

9. Hadoop

1. Пример благоразумного использования Hadoop

i. Анализ 10 Гб данных.

ii. Ежедневное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).

iii. Посекундное сохранение данных температуры, поступающих со всех городов России (по одному показанию на город, всего городов 1100 шт).

iv. Построение графика пульса пациента в реальном времени.

2. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?

i. 100Гб

ii. 1Тб

iii. 100Тб

iv. 1Пб

3. Hadoop – это:

i. Набор утилит, и программный каркас для выполнения распределённых программ, работающих на кластерах.

ii. Распределённая СУБД, позволяющая обрабатывать большие данные.

iii. Язык выполнения заданий в парадигме MapReduce.

iv. Распределённая файловая система, предназначенная для хранения файлов большого объёма.

1. Перечислите четыре основных характеристики Big Data:

1. Virtualization, Volume, Variability, Vehicle
2. **Variety, Velocity, Volume, Value**
3. Verification, Volume, Velocity, Visualization
4. Video, Value, Variety, Volume

Ответ: 2

2.Укажите соответствие

Левая часть	**Правая часть**
1. Структурированные данные	a) Библиотека фильмов, представленных в формате mpeg4 на одном жестком диске
2. Полу-структурированные данные	b)Таблица с ежедневными наблюдениями температуры в формате csv
3. Не структурированные данные	3. Текст дипломной работы, представленный в формате PDF
	4. данные о продажах компании, представленные в виде помесячных отчётов в формате MS Word
	5. Файл данными в формате XML
	6. Файл с данными в формате exel

Ответы: 1-b, 1-f, 2-e, 3-a, 3-d, 3-c

3. Какие из задач решаются Big Data?

(1) Мониторинг оборудования
(2) Анализ социальных сетей
(3) Оптимизация автомобильного движения
(4) Все вышеперечисленное

Ответ: 4

4, Какие существуют виды грязных данных (множественный выбор):

1. пропущенные значения;
2. дубликаты данных;
3. нулевые значения
4. шумы и выбросы.

Ответ: 1, 2, 4.

5.Какие методы можно использовать для заполнения пропущенных данных (множественный выбор):

1. удаления каждой строки, содержащей пропущенные значения
2. заполнение средним значением
3. заполнение медианой
4. заполнение модой
5. дублирование соседней строки

Ответ: 1,2.3.4.

Темы контрольных заданий

1. Метод наименьших квадратов применительно к задаче линейной регрессии.

2. Логистическая регрессия.

3. Наивный классификатор Байеса.

4. Алгоритм k-means.

5. Алгоритм Априори.

6. Использование готовых решений анализа данных (Longinom, RapidMiner, Weka и т.д.).

7. Визуализация данных .

8. Алгоритм k-means, реализация в рамках парадигмы Map Reduce.

9. Регуляризация метода наименьших квадратов.

11. Алгоритм SVM.

20.2 Промежуточная аттестация

Промежуточная аттестация может включать в себя проверку теоретических вопросов, а также, при необходимости (в случае не выполнения в течение семестра), проверку выполнения установленного перечня лабораторных заданий, позволяющих оценить уровень полученных знаний и/или практическое (ие) задание(я), позволяющее (ие) оценить степень сформированности умений и навыков.

Для оценки теоретических знаний используется перечень контрольно-измерительных материалов. Каждый контрольно-измерительный материал для проведения промежуточной аттестации включает два задания - вопросов для контроля знаний, умений и владений в рамках оценки уровня сформированности компетенции. При оценивании используется количественная шкала. Критерии оценивания представлены в приведенной ниже таблице

Для оценивания результатов обучения на экзамене используются следующие содержательные показатели (формулируется с учетом конкретных требований дисциплины)

1. знание теоретических основ учебного материала, основных определений, понятий и используемой терминологии;

2. владение навыками проведения компьютерного эксперимента, тестирования компьютерных алгоритмов обработки информации.

3. владение навыками программирования и экспериментирования с компьютерными моделями алгоритмов обработки информации в рамках выполняемых лабораторных заданий;

4. умение обосновывать свои суждения и профессиональную позицию по излагаемому вопросу;

5. умение связывать теорию с практикой, иллюстрировать ответ примерами, в том числе, собственными, умение выявлять и анализировать основные закономерности, полученные, в том числе, в ходе выполнения лабораторно-практических заданий;

6. умение проводить обоснование и представление основных теоретических и практических результатов (теорем, алгоритмов, методик) с использованием математических выкладок, блок-схем, структурных схем и стандартных описаний к ним;

Различные комбинации перечисленных показателей определяют критерии оценивания результатов обучения (сформированности компетенций) на зачете: высокий (углубленный) уровень сформированности компетенций; повышенный (продвинутый) уровень сформированности компетенций; пороговый (базовый) уровень сформированности компетенций.Для оценивания результатов обучения на зачете с оценкой используется 4-балльная шкала: «отлично», «хорошо», «удовлетворительно», «неудовлетворительно». Для оценивания результатов обучения на зачете используется – зачтено, не зачтено по результатам тестирования. Соотношение показателей, критериев и шкалы оценивания результатов обучения на зачете с оценкой представлено в следующей таблице.

Критерии оценивания компетенций	Уровень сформированности компетенций	Шкала оценок
Обучающийся демонстрирует полное соответствие знаний, умений, навыков по приведенным критериям свободно оперирует понятийным аппаратом и приобретенными знаниями, умениями, применяет их при решении практических задач. Успешно выполнены лабораторные работы в соответствии с установленным перечнем.	Повышенный уровень	Отлично
Ответ на контрольно-измерительный материал не полностью соответствует одному из перечисленных выше показателей, но обучающийся дает правильные ответы на дополнительные вопросы. При этом обучающийся демонстрирует соответствие знаний, умений, навыков приведенным в таблицах показателям, но допускает незначительные ошибки, неточности, испытывает затруднения при решении практических задач. Успешно выполнены лабораторные работы в соответствии с установленным перечнем.	Базовый уровень	Хорошо
Обучающийся демонстрирует неполное соответствие знаний, умений, навыков приведенным в таблицах показателям, допускает значительные ошибки при решении практических задач. При этом ответ на контрольно-измерительный материал не соответствует любым двум из перечисленных показателей, обучающийся дает неполные ответы на дополнительные вопросы. Успешно выполнены лабораторные работы в соответствии с установленным перечнем.	Пороговый уровень	Удовлетворительно

Ответ на контрольно-измерительный материал не соответствует любым трем из перечисленных показателей. Обучающийся демонстрирует отрывочные, фрагментарные знания, допускает грубые ошибки. Не выполнены лабораторные работы в соответствии с установленным перечнем.	–	Неудовлетворительно

УТВЕРЖДАЮ

Заведующий кафедрой технологий обработки и защиты информации

_____ А.А. Сирота

__.__.2024

Направление подготовки / специальность 09.04.02 Информационные системы и технологии

Дисциплина Б1.В.05 Анализ больших данных

Форма обучения Очное

Вид контроля зачет с оценкой

Вид аттестации Промежуточная

Контрольно-измерительный материал № 1

1. Средства массово-параллельной обработки неопределённо структурированных данных - решениями категории NoSQL, алгоритмы MapReduce, программные каркасы и библиотеки проекта Hadoop.
2. Методы и техники анализа, применимые к большим данным: классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным).
3. Практическое задание (проведение предобработки и очиски данных на языке Python).

Преподаватель _____ В.В.Гаршина

№	Вопросы по курсу
1	Большие данные (big data) в информационных технологиях. Понятие Больших данных. Особенности сбора, хранения, обработки и анализа больших массивов данных.
2	Три V: объём (volume), скорость (velocity), многообразие (variety). Источники больших данных. Использование больших данных в науке, бизнесе, государственном управлении.
3	Средства массово-параллельной обработки неопределённо структурированных данных - решениями категории NoSQL, алгоритмы MapReduce, программные каркасы и библиотеки проекта Hadoop.
4	Методы и техники анализа, применимые к большим данным: методы класса Data Mining: обучение ассоциативным правилам (association rule learning),
5	Методы и техники анализа, применимые к большим данным: классификация (методы категоризации новых данных на основе принципов, ранее применённых к уже наличествующим данным).

6	Методы и техники анализа, применимые к большим данным: методы класса Data Mining: кластерный анализ, регрессионный анализ.
7	Техники, интеграции разнородных данных из разнообразных источников для возможности глубинного анализа. Краудсорсинг, принципы смешения и интеграции данных.
8	Использование моделей, построенных на базе статистического анализа или машинного обучения для получения комплексных прогнозов.
9	Аналитика; имитационное моделирование.
10	Пространственный анализ (англ. Spatial analysis) - класс методов, использующих топологическую, геометрическую и географическую информацию в данных
11	Цели и возможности визуализации аналитических данных - представление информации в виде рисунков, диаграмм и анимации.
12	Понятие прогноза. Закон распределения случайной величины. Статистические оценки параметров. Доверительные области.
13	Теория моментов. Корреляционный анализ.
14	Использование модели множественной линейной регрессии для прогнозирования показателей.
15	Сглаживание временных рядов. Динамические модели с распределенными лагами.
16	Стационарные временные ряды. Тестирование стационарности.
17	Коинтеграция. Анализ временных рядов.
18	Экспоненциальное сглаживание.
19	Авторегрессионные модели.
20	Идентификация авторегрессионной модели скользящего среднего.
21	Прогнозирование с моделями временных рядов. Доверительные интервалы прогноза.
22	Дисперсионный анализ влияния качественных факторов. Ранговые методы.
23	Факторный анализ. Метод главных факторов.
24	Классическая модель многомерного шкалирования.
25	Кластерный анализ.
26	Дискриминантный анализ.
27	Многомерный статистический анализ.
28	Требования к распределенным информационным системам. Средства построения распределенных информационных систем
29	Технология Map-Reduce
30	Система Apache Hadoop
31	Базы данных NoSQL. Особенности, классификация.

32	Возможности NoSQL-баз данных по обеспечению целостности, доступности скорости обработки информации. CAP-теорема.
33	Документо-ориентированные базы данных.
34	Возможности СУБД MongoDB
35	Работа с документо-ориентированными БД на языке JSON.