МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Воронежский государственный технический университет»

УТВЕРЖДАЮ

И.о. декана факультета информационных технологий и компьютерной безопасности

_____ / А.В. Бредихин /

19.03.2024

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Технологии обработки больших данных»

Направление подготовки 09.03.02 Информационные системы и технологии

Профиль Технологии искусственного интеллекта

Квалификация выпускника бакалавр

Нормативный период обучения 4 года

Форма обучения очная

Год начала подготовки 2024

Автор программы _____ С.Ю. Белецкая

И.о. заведующего кафедрой систем автоматизированного проектирования и информационных систем _____ П.Ю. Гусев

Руководитель ОПОП _____ Д.В. Иванов

Воронеж 2024

1. ЦЕЛИ И ЗАДАЧИ ДИСЦИПЛИНЫ

1.1. Цели дисциплины

Целью дисциплины является освоение принципов, методов, технологий и инструментов хранения и обработки больших данных

1.2. Задачи освоения дисциплины

- ознакомление студентов с основными классами задач обработки больших данных;
- изучение методов и технологий интеллектуального анализа больших объёмов данных;
- освоение технологий подготовки, хранения и распределённой обработки больших данных;
- приобретение навыков использования технологий и инструментов Big Data при решении практических задач.

2. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОПОП

Дисциплина «Технологии обработки больших данных» относится к дисциплинам части, формируемой участниками образовательных отношений блока Б1.

3. ПЕРЕЧЕНЬ ПЛАНИРУЕМЫХ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ ПО ДИСЦИПЛИНЕ

Процесс изучения дисциплины «Технологии обработки больших данных» направлен на формирование следующих компетенций:

ПК-3 - Способен осуществлять сбор и подготовку данных, анализировать структуры данных, проектировать и разрабатывать базы данных в рамках создания (модификации) и сопровождения информационных систем

ПК-5 - Способен осуществлять сбор и подготовку данных для систем искусственного интеллекта

ПК-8 - Способен выявить естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности в области моделирования и анализа сложных естественных и искусственных систем

Компетенция	Результаты обучения, характеризующие сформированность компетенции
ПК-3	**Знать** основные особенности и источники больших данных, технологии хранения и анализа больших данных
	Уметь проектировать и разрабатывать нереляционные базы данных
	Владеть навыками хранения и распределённой обработки больших данных при разработке и эксплуатации информационных систем
ПК-5	**Знать** технологии сбора и подготовки данных для систем искусственного интеллекта
	Уметь осуществлять сбор данных из разных источников

	Владеть инструментальными средствами обработки больших данных
ПК-8	**Знать** модели и методы интеллектуального анализа больших данных
	Уметь использовать технологию интеллектуального анализа данных в процессе формализации и алгоритмизации профессиональных задач
	Владеть навыками разработки программного обеспечения для решения аналитических задач обработки информации

4. ОБЪЕМ ДИСЦИПЛИНЫ

Общая трудоемкость дисциплины «Технологии обработки больших данных» составляет 6 з.е.

Распределение трудоемкости дисциплины по видам занятий

очная форма обучения

Виды учебной работы	Всего часов	Семестры	
		7	8
Аудиторные занятия (всего)	84	54	30
В том числе:			
Лекции	28	18	10
Лабораторные работы (ЛР)	56	36	20
Самостоятельная работа	96	54	42
Курсовой проект	+		+
Часы на контроль	36		36
Виды промежуточной аттестации экзамен зачёт	 + +	 +	 +
Общая трудоемкость: академические часы зач.ед.	 216 6	 108 3	 108 3

5. СОДЕРЖАНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

5.1 Содержание разделов дисциплины и распределение трудоемкости по видам занятий

очная форма обучения

№ п/п	Наименование темы	Содержание раздела	Лекц	Лаб. зан.	СРС	Всего, час
1	Основы Big Data	Понятие больших данных. Особенности сбора, хранения, обработки и анализа больших массивов данных. Источники больших данных. Характеристики больших данных. Жизненный цикл больших данных. Этапы обработки больших данных. Задачи, решаемые при обработке больших данных. Принципы управления большими дан-	4	4	16	24

		ными. Требования к системам хранения и обработки больших данных. Использование больших данных в информационных системах				
2	Аналитика больших данных	Основные задачи анализа больших данных в информационных системах. Классификация современных методов анализа данных. Разведочный анализ больших данных. Сбор данных из разных источников. Предварительный анализ, обработка и очистка данных. Особенности технологии Data Mining. Задачи Data Mining и методы их решения. Использование методов Data Mining для анализа больших данных. Решение задач классификации, кластеризации, прогнозирования на основе больших данных. Методы машинного обучения, их классификация и особенности. Использование методов Machine learning для анализа больших данных. Нейросетевые технологии обработки информации. Технологии визуализации и трансформации данных. Многомерный анализ данных.. Инструментарий анализа больших данных.	6	12	25	43
3	Технологии и инструменты распределённой обработки больших данных	Базовые архитектуры распределённой обработки данных. Параллельные и облачные вычисления. Модель распределённых вычислений MapReduce, ей назначение, особенности, области применения. Принцип работы MapReduce. Этапы обработки данных на основе MapReduce, базовые функции вычислительной модели. Платформа распределённой обработки больших данных Hadoop. Возможности, основы построения и архитектура Hadoop. Элементы кластера Hadoop. Основные компоненты Hadoop. Основы Yarn. Реализация и использование фреймворка MapReduce в распределённой среде. Распределённая файловая система HDFS, архитектура, принципы функционирования и реализация. Языки поисковых запросов для Hadoop. Состав экосистемы Hadoop, основные проекты и технологии. Apache Spark, его особенности, архитектура, принципы функционирования. Организация распределённой обработки слабоструктурированных данных на основе Apache Spark. Распределённая потоковая платформа Apache Kafka, возможности, архитектура, принципы функционирования. Организация потоковой передачи данных на основе Apache Kafka. Основные принципы работы Apache Hive. Выполнение запросов, агрегирование и анализ данных. Использование современных инструментов BigData при проектировании и эксплуатации информационных систем.	8	20	30	58
4	Технологии хранения больших данных. Организа-	Концепции хранилищ данных. Требования к хранилищам данных. Архитектура и принципы построения хранилищ данных.	10	20	25	55

	ция хранилищ данных на основе NoSQL СУБД	Информационные потоки в хранилище данных. Технологии хранения больших данных на основе нереляционных СУБД. Основные характеристики NoSQL СУБД. Особенности нереляционных СУБД. Ключевые отличия между реляционными и нереляционными СУБД. Модели данных в нереляционных СУБД. Формулировка CAP теоремы и её следствий. Классификация NoSQL СУБД. Базы данных “ключ-значение”. Ассоциативные массивы, кэш-память и алгоритмы вытеснения. Принципы построения баз данных “ключ -значение. Прикладное использование. Примеры. Принципы организации СУБД Redis. Документо ориентированные базы данных. Принципы организации документо-ориентированных баз данных. Области использования. Пример ДОСУБД MongoDB. Запросы к СУБД на языке JSON. Программные интерфейсы для работы с MongoDB. Принципы построения колоночных СУБД. Модели представления данных, организация запросов. Примеры колоночных СУБД – Hbase, Cassandra. Графовые СУБД, их особенности и области применения. Способы организации и хранения аннотированных графов. Характеристики графовых баз данных. Графовая СУБД Neo4j. Программные интерфейсы для работы с Neo4j. Использование нереляционных СУБД при проектировании и эксплуатации информационных систем.				
		Итого	**28**	**56**	**96**	**180**

5.2 Перечень лабораторных работ

1-2. Обработка и анализ данных с использованием библиотек Python

3. Платформа Hadoop и ее компоненты

4. Распределенные файловая система HDFS

5. Модель вычислений MapReduce

6-7. Распределенные вычисления на платформе Apache Spark

8. Apache Kafka. Потоковая обработка данных

9. Apache Hive. Выполнение запросов, агрегирование и анализ данных.

10. Изучение возможностей и работа с документо-ориентированной СУБД MongoDB.

11. Работа с NoSQL СУБД колоночного типа HBase.

12. Работа с NoSQL СУБД Redis. Проектирование модели Key Value для Redis.

6. ПРИМЕРНАЯ ТЕМАТИКА КУРСОВЫХ ПРОЕКТОВ (РАБОТ) И КОНТРОЛЬНЫХ РАБОТ

В соответствии с учебным планом освоение дисциплины предусматривает выполнение курсового проекта в 8 семестре для очной формы обучения.

Примерная тематика курсового проекта: «Создание NoSQL базы данных в конкретной предметной области и разработка приложения»

Задачи, решаемые при выполнении курсового проекта:

- сформулировать цель создания базы данных;
- описать возможного пользователя базы данных;
- определить круг запросов и задач, которые предполагается решать с использованием созданной базы данных;
- сформулировать требования к базе данных;
- построить модель;
- осуществить и обосновать выбор СУБД и технических средств;
- создать базу данных с использованием выбранной СУБД;
- разработать приложение для реализации запросов и решения задач.

Курсовой проект включат в себя разработанное приложение и расчетно-пояснительную записку.

7. ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ДЛЯ ПРОВЕДЕНИЯ ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ ОБУЧАЮЩИХСЯ ПО ДИСЦИПЛИНЕ

7.1. Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

7.1.1 Этап текущего контроля

Результаты текущего контроля знаний и межсессионной аттестации оцениваются по следующей системе:

«аттестован»;

«не аттестован».

Компетенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Аттестован	Не аттестован
ПК-3	**Знать** основные особенности и источники больших данных, технологии хранения и анализа больших данных	Знание основных подходов к хранению больших данных, принципов организации и структуры хранилищ данных, особенностей и основных классов NoSQL СУБД. Ответы на теоретические вопросы при защите лабораторных работ и курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Уметь проектировать и разрабатывать нереляционные базы данных	Умение проектировать модели данных, разрабатывать нереляционные базы для хранения больших данных. Вы-	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах

		полнение лабораторных работ и курсового проекта		
	Владеть навыками хранения и распределённой обработки больших данных при разработке и эксплуатации информационных систем	Владение навыками разработки баз данных в различных предметных областях на основе NoSQL СУБД. Выполнение лабораторных работ и курсового проекта.	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ПК-5	**Знать** технологии сбора и подготовки данных для систем искусственного интеллекта	Знание основных этапов обработки данных, методов и средств сбора и подготовки данных для систем искусственного интеллекта Ответы на теоретические вопросы при защите лабораторных работ и курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Уметь осуществлять сбор данных из разных источников	Умение анализировать источники больших данных, осуществлять сбор и агрегирование данных из разных источников. Умение проводить подготовку данных для их дальнейшего использования в системах искусственного интеллекта. Выполнение лабораторных работ и курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владеть инструментальными средствами обработки больших данных	Владение навыками использования инструментальных средств Big Data при решении практических задач. Выполнение лабораторных работ и курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
ПК-8	**Знать** модели и методы интеллектуального анализа больших данных	Знание методов и средств интеллектуального анализа больших данных. Ответы на теоретические вопросы при защите лабораторных работ и курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Уметь использовать технологию интеллектуального анализа данных в процессе формализации и алгоритмизации профессиональных задач	Умение решать задачи анализа и распределённой обработки больших данных в различных предметных областях. Выполнение лабораторных работ и курсового проекта	Выполнение работ в срок, предусмотренный в рабочих программах	Невыполнение работ в срок, предусмотренный в рабочих программах
	Владеть навыками разработки про-	Владение навыками разработки программ-	Выполнение работ в срок, предусмотренный в рабо-	Невыполнение работ в срок, предусмотрен-

	граммного обеспечения для решения аналитических задач обработки информации	ного обеспечения для решения аналитических задач обработки больших данных	чих программах	ный в рабочих программах

7.1.2 Этап промежуточного контроля знаний

Результаты промежуточного контроля знаний оцениваются в 7, 8 семестре для очной формы обучения по двух/четырехбальной системе:

«зачтено»
«не зачтено» (для зачёта в 7 семестре)

Компетенция	Результаты обучения, характеризующие сформированность компетенции	Критерии оценивания	Зачтено	Не зачтено
ПК-3	**Знать** основные особенности и источники больших данных, технологии хранения и анализа больших данных	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь проектировать и разрабатывать нереляционные базы данных	Решение стандартных практических задач	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками хранения и распределённой обработки больших данных при разработке и эксплуатации информационных систем	Решение прикладных задач в конкретной предметной области	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
ПК-5	**Знать** технологии сбора и подготовки данных для систем искусственного интеллекта	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь осуществлять сбор данных из разных источников	Решение стандартных практических задач	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеть инструментальными средствами обработки больших данных	Решение прикладных задач в конкретной предметной области	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
ПК-8	**Знать** модели и методы интеллектуального анализа больших данных	Тест	Выполнение теста на 70-100%	Выполнение менее 70%
	Уметь использовать технологию интеллектуального анализа данных в процессе формализации и алгоритмизации про-	Решение стандартных практических задач	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены

	фессиональных задач			
	Владеть навыками разработки программного обеспечения для решения аналитических задач обработки информации	Решение прикладных задач в конкретной предметной области	Продемонстрирован нверный ход решения в большинстве задач	Задачи не решены

или
«отлично»;
«хорошо»;
«удовлетворительно»;
«неудовлетворительно» (для экзамена в 8 семестре)

Компетенция	**Результаты обучения, характеризующие сформированность компетенции**	**Критерии оценивания**	**Отлично**	**Хорошо**	**Удовл.**	**Неудовл.**
ПК-3	**Знать** основные особенности и источники больших данных, технологии хранения и анализа больших данных	Тест	Выполнение теста на 90- 100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов
	Уметь проектировать и разрабатывать нереляционные базы данных	Решение стандартных практических задач	Задачи решены в полном объеме и получены верные ответы	Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками хранения и распределённой обработки больших данных при разработке и эксплуатации информационных систем	Решение прикладных задач в конкретной предметной области	Задачи решены в полном объеме и получены верные ответы	Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
ПК-5	**Знать** технологии сбора и подготовки данных для систем искусственного интеллекта	Тест	Выполнение теста на 90- 100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов

	Уметь осуществлять сбор данных из разных источников	Решение стандартных практических задач	Задачи решены в полном объеме и получены верные ответы	Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеть инструментальными средствами обработки больших данных	Решение прикладных задач в конкретной предметной области	Задачи решены в полном объеме и получены верные ответы	Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
ПК-8	**Знать** модели и методы интеллектуального анализа больших данных	Тест	Выполнение теста на 90- 100%	Выполнение теста на 80-90%	Выполнение теста на 70-80%	В тесте менее 70% правильных ответов
	Уметь использовать технологию интеллектуального анализа данных в процессе формализации и алгоритмизации профессиональных задач	Решение стандартных практических задач	Задачи решены в полном объеме и получены верные ответы	Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены
	Владеть навыками разработки программного обеспечения для решения аналитических задач обработки информации	Решение прикладных задач в конкретной предметной области	Задачи решены в полном объеме и получены верные ответы	Продемонстрирован верный ход решения всех, но не получен верный ответ во всех задачах	Продемонстрирован верный ход решения в большинстве задач	Задачи не решены

7.2 Примерный перечень оценочных средств (типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности)

7.2.1 Примерный перечень заданий для подготовки к тестированию

1 Какими свойствами, как правило, обладают большие данные? (возможно несколько ответов)

а) структурированность;

б) неструктурированность;

в) объем в терабайтах и петабайтах;

г) объем в мегабайтах и гигабайтах.

2 Какая технология используется для оперативного анализа данных?

а) Data Mining;

б) OLTP;

в) DSS;

г) OLAP.

3. Какие модели хранения данных поддерживают базы данных NoSQL? (возможно несколько ответов)

а) столбчатые;
б) ключ-значение;
в) графовые;
г) реляционные.

4 Какой метод применяется для обнаружения скрытых знаний в больших объемах данных?

а) **Data Mining;**
б) дискриминантный анализ;
в) OLAP;
г) дисперсионный анализ.

5 Какие методы применяются для анализа неструктурированных текстовых данных? (возможно несколько ответов)

а) DSS Mining;
б) Web Mining;
в) Machine Mining;
г) Text Mining.

6. Какой метод используется для разделения элементов выборки на множество групп в зависимости от выявленных внутренних связей между ними?

факторный анализ;
дисперсионный анализ;
дискриминантный анализ;
кластерный анализ.

7. Какая базовая структура данных стоит в основе модели MapReduce?

а) пара (ключ, значение)
б) бинарное дерево
в) массив произвольной длины
г) функция высшего порядка

8 Какая из баз данных работает на основе стека Hadoop?

а) HBase
б) Dynamo
в Bigtable
г) Cassandra

9 Для хранения частей базы данных на разных серверах применяется:

а**) шардинг**
б) денормализация
в) репликация
г) унификация

10. HDFS – это…
а**)** СУБД
б) среда программирования
в) распределённая файловая система
г) система анализа данных

7.2.2 Примерный перечень заданий для решения стандартных задач

1 Какая из библиотек Python предназначена для работы с нейронными сетями?

Pandas
Keras
NumPy
Statsmodels

2 Какая из библиотек Python предназначена для машинного обучения?

Scikit Learn
Scipy
NumPy
Pandas

3. В чем заключаются отличия задач классификации и кластеризации данных?

отличий нет, эти понятия являются синонимами;
при кластеризации заранее известны группы, к которым должен быть отнесен объект;
при классификации заранее известны группы, к которым должен быть отнесен объект;
при классификации заранее не известны группы, к которым должен быть отнесен объект.

4 Какая из перечисленных систем не является компонентом Hadoop:
а**)** MapReduce
б) Yarn
в) HDFS

г) SPSS

5 Какая из перечисленных NoSQL СУБД относится к документоориентированному типу?

1) HBase
2) MongoDB
3) Redis
4) Cassandra

6 Выберите верные утверждения о модели данных в HBase.

1) внутри ячеек таблицы поддерживается хранение нескольких версий данных
2) версионность данных не поддерживается на уровне стандартных средств
3) количество колонок созданной таблицы по умолчанию является наперед заданным
4) колонки семейства можно идентифицировать по их общему префиксу

7 Глубина вложенности документов в MobgoDB:

1) **не ограничена**
2) ограничена 2-мя уровнями вложенности
3) ограничена 1-им уровнем вложенности
4) ограничена 3-мя уровнями вложенности

8 Какая из перечисленных NoSQL СУБД относится к колоночному типу?

1) Neo4j
2) MongoDB
3) Redis
4) Cassandra

9 Опишите работу функции map.

1) функция map возвращает некоторую функцию с аргументом в виде данного списка
2) функция map объединяет поэлементно два и более списка, возвращая список пар элементов
3) функция map осуществляет поиск данного элемента в списке-аргументе
4) функция map применяет данную функцию к каждому элементу списка, возвращая список результатов

10 На каких узлах в Hadoop MapReduce выполняются map- и reduce-задачи?

1) master nodes
2) job submission nodes
3) slave nodes

4) namenodes

11.Каков размер блока HDFS по умолчанию (для релиза 2.6.0 и выше)?

1) 512 байт
2) **128 Мбайт**
3) 256 Мбайт
4) 64 Мбайт

7.2.3 Примерный перечень заданий для решения прикладных задач

1 Какой из методов Data Mining может использоваться для определения групп потребителей со схожими стереотипами поведения с целью предложения интересных именно им товаров?

а) кластерный анализ;
б) генетический алгоритм;
в) поиск ассоциативных правил;
г) нечеткая логика.

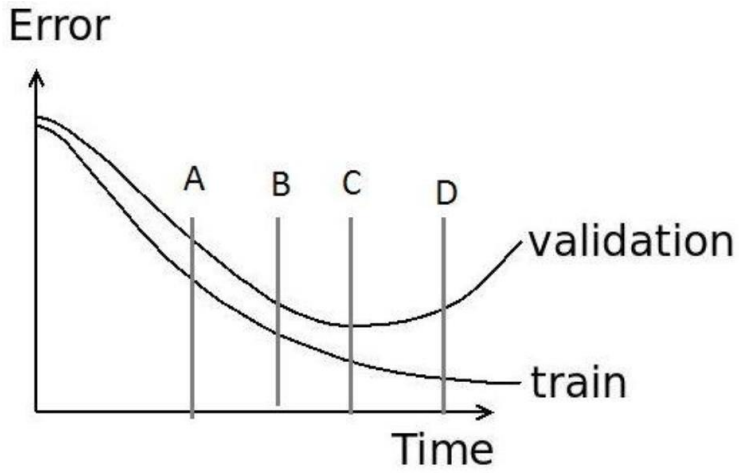
2 Имеются данные, характеризующие уровень инфляции в регионе:

Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь
2,99	2,66	2,63	2,56	2,40	2,22	1,97	1,72	1,56	1,42

Используя метод наименьших квадратов (линейная функция), определить прогноз уровня инфляции в декабре.

1)1,35
2)1,11
3)1,43
4)1,24

3 При обучении нейронной сети в задаче распознавания образов был построен следующий график ошибки обучения и ошибки валидации модели. Какой момент времени лучше всего подходит для раннего останова?



а) A
б) B
в) C
г) D

4. Опишите структуру нейронной сети, созданной с помощью программного кода:

```
model = Sequential([Dense(64, input_dim=100), activation('relu'),
Dense(10), activation('softmax'),])
```

1) MLP 100-64-10, функция активации скрытого слоя relu, функция активации выходного слоя softmax

2) MLP 100-10-64, функция активации скрытого слоя relu, функция активации выходного слоя softmax

3) MLP 100-64-10, функция активации скрытого слоя softmax, функция активации выходного слоя relu

4) MLP 100-10-64, функция активации скрытого слоя relu, функция активации выходного слоя softmax

5 Как следует отредактировать запрос db.teachers.update({name:'Василий'}, {$inc: {salary:5000}}), чтобы обновить все найденные документы?

1) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},false,false,true)
2) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},false)
3) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},false,true)
4) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},true)

6. Результатом выполнения операции db.unicorns.find().skip(5) в MongoDB будет:

а) возврат первых 5 документов

б) возврат последних 5 документов
в) возврат всех документов, содержащих цифру 5
г) пропуск первых 5 документов коллекции и возврат всех остальных документов

7. Как следует отредактировать запрос db.teachers.update({name:'Василий'}, {$inc: {salary:5000}}), чтобы разрешить вставку документа, если его не существует в коллекции?

1) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},false,false,true)
2) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},false)
3) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},false,true)
4) db.teachers.update({name:'Василий'}, {$inc: {salary:5000}},true)

7 Для чего предназначена утилита distcp?

1) Для параллельного копирования больших объёмов данных из локальной файловой системы в HDFS
2) Для параллельного копирования образа файловой системы между основным узлом имён и резервным
3) Для параллельного копирования больших объёмов данных между кластерами Hadoop
4) Ни один из указанных вариантов

8 Для чего предназначена команда hdfs dfsadmin -refreshNodes?

1) Проверяет доступность всех узлов данных кластера и выдаёт список существующих, но недоступных по каким-либо причинам узлов
2) Вызывает повторное чтение конфигурационных файлов на всех узлах кластера (NameNode и DataNode)
3) Обновляет узел имён (NameNode) новым набором разрешённых узлов данных
4) Повторно запускает фоновые процессы на всех узлах кластера

9. Для агрегации данных в случаях отношения «один-ко-многим» или «многие-ко- многим» классические СУРБД применяют конструкцию JOIN. Для таких случаев в MongoDB используется

а) аналог конструкции JOIN
б) только вложенные документы
в) только технология DBRef
г) вложенные документы и технология DBRef

10 Результатом выполнения команды db.source.copyTo(target) будет

1) **копирование всех документов из одной коллекции в другую**
2) копирование всех полей из одного документа в другой
3) копирование определенных полей документа
4) копирование всех коллекций

11 Какое количество фоновых процессов менеджера ресурсов и менеджера узлов работает при стандартной конфигурации кластера с YARN архитектурой?

1**) Один процесс менеджера ресурсов и один или более процессов менеджера узлов (по одному на каждом узле кластера)**

2) Один процесс менеджера ресурсов и один процесс менеджера узлов

3) Один или более процессов менеджера ресурсов (по одному на каждом узле кластера) и один процесс менеджера узлов

4) Один или более процессов менеджера ресурсов (по одному на каждом узле кластера) и один или более процессов менеджера узлов (по одному на каждом узле кластера)

7.2.4 Примерный перечень вопросов для подготовки к зачету

1 Понятие больших данных. Основные задачи обработки больших данных в информационных системах.

2 Характеристики больших данных.

3 Жизненный цикл больших данных. Этапы обработки больших данных.

4 Принципы управления большими данными. Требования к системам хранения и обработки больших данных.

5. Основные задачи анализа данных в интеллектуальных информационных системах.

7. Источники больших данных. Технологии сбора данных из разных источников.

8. Классификация современных методов анализа данных. Понятие о технологии Data Mining.

9. Этапы интеллектуального анализа данных.

10. Первичный разведочный анализ данных.

11. Технология предварительной обработки данных в интеллектуальных системах.

12. Решение задач кластеризации на основе больших данных.

13. Решение задач классификации на основе больших данных.

14. Решение задач прогнозирования на основе больших данных.

15. Методы машинного обучения, их классификация и особенности.

16. Нейросетевые технологии анализа данных.

17. Технологии визуализации и трансформации данных. Многомерный анализ данных. OLAP-технологии.

18. Инструментарий анализа больших данных

19. Требования к распределенным и системам обработки больших данных.
20. Решение задач трансформации и визуализации данных.

7.2.5 Примерный перечень заданий для подготовки к экзамену

1 Модель распределённых вычислений MapReduce, ей назначение, особенности, области применения.

2 Принцип работы MapReduce. Этапы обработки данных на основе MapReduce, базовые функции вычислительной модели.

3. Платформа распределённой обработки больших данных Hadoop. Возможности, основы построения и архитектура Hadoop. Основные компоненты Hadoop.

4. Реализация и использование фреймворка MapReduce в распределённой среде.

5. Распределённая файловая система HDFS, архитектура, принципы функционирования и реализация. Языки поисковых запросов для Hadoop.

6. Состав экосистемы Hadoop, основные проекты и технологии.

7. Apache Spark, его особенности, архитектура, принципы функционирования. Организация распределённой обработки слабоструктурированных данных на основе Apache Spark.

8. Распределённая потоковая платформа Apache Kafka, возможности, архитектура, принципы функционирования. Организация потоковой передачи данных на основе Apache Kafka.

9. Основные принципы работы Apache Hive. Выполнение запросов, агрегирование и анализ данных.

10. Использование современных инструментов BigData при проектировании и эксплуатации информационных систем.

11. Базы данных NoSQL. Особенности, классификация

12 Возможности NoSQL СУБД по обеспечению целостности, доступности скорости обработки информации. CAP-теорема.

13 Документо-ориентированные СУБД

14 Возможности СУБД MongoDB. Модели данных. Организация запросов.

15. Базы данных “Ключ-значение”. Принципы построения, примеры.

16. Колоночные СУБД. Принципы построения, примеры.

17. Графовые СУБД. Принципы построения, области применения, примеры.

18. Использование нереляционных СУБД при проектировании и эксплуатации информационных систем.

7.2.6. Методика выставления оценки при проведении промежуточной аттестации

Зачёт в 7 семестре проводится по тест-билетам, каждый из которых содержит 10 вопросов и задачу. Каждый правильный ответ на вопрос в тесте

оценивается 1 баллом, задача оценивается в 10 баллов (5 баллов верное решение и 5 баллов за верный ответ). Максимальное количество набранных баллов – 20.

1. Зачёт ставится в случае, если студент набрал более 10 баллов.
2. Незачёт ставится в случае, если студент набрал менее 10 баллов

Экзамен в 8 семестре проводится по тест-билетам, каждый из которых содержит 10 вопросов и задачу. Каждый правильный ответ на вопрос в тесте оценивается 1 баллом, задача оценивается в 10 баллов (5 баллов верное решение и 5 баллов за верный ответ). Максимальное количество набранных баллов – 20.

1. Оценка «Неудовлетворительно» ставится в случае, если студент набрал менее 6 баллов.

2. Оценка «Удовлетворительно» ставится в случае, если студент набрал от 6 до 10 баллов

3. Оценка «Хорошо» ставится в случае, если студент набрал от 11 до 15 баллов.

4. Оценка «Отлично» ставится, если студент набрал от 16 до 20 баллов

7.2.7 Паспорт оценочных материалов

№ п/п	Контролируемые разделы (темы) дисциплины	Код контролируемой компетенции	Наименование оценочного средства
1	Основы Big Data	ПК-3, ПК-5, ПЕ-8	Тест, защита лабораторных работ, требования к курсовому проекту
2	Аналитика больших данных	ПК-3, ПК-5, ПК-8	Тест, защита лабораторных работ, требования к курсовому проекту
3	Технологии и инструменты распределённой обработки больших данных	ПК-3, ПК-5, ПК-8	Тест, защита лабораторных работ, требования к курсовому проекту
4	Технологии хранения больших данных. Организация хранилищ данных на основе NoSQL СУБД	ПК-3, ПК-5, ПК-8	Тест, защита лабораторных работ, требования к курсовому проекту

7.3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Тестирование осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных тест-заданий на бумажном носителе. Время тестирования 30 мин. Затем осуществляется проверка теста экзаменатором и выставляется оценка согласно методики выставления оценки при проведении промежуточной аттестации.

Решение стандартных задач осуществляется, либо при помощи компьютерной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

Решение прикладных задач осуществляется, либо при помощи компью-

терной системы тестирования, либо с использованием выданных задач на бумажном носителе. Время решения задач 30 мин. Затем осуществляется проверка решения задач экзаменатором и выставляется оценка, согласно методики выставления оценки при проведении промежуточной аттестации.

Защита курсового проекта осуществляется согласно требованиям, предъявляемым к работе, описанным в методических материалах. Примерное время защиты на одного студента составляет 20 мин.

8 УЧЕБНО МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ)

8.1 Перечень учебной литературы, необходимой для освоения дисциплины

1 Черемухин, А. Д. Большие данные [Электронный ресурс]:: учебное пособие / А. Д. Черемухин. — Москва : Ай Пи Ар Медиа, 2023. — 782 с. —Режим доступа: https://www.iprbookshop.ru/129721

2 Чубукова, И. А. Data Mining : учебное пособие / И. А. Чубукова. — 4-е изд. — Москва : Интернет-Университет Информационных Технологий (ИНТУИТ), Ай Пи Ар Медиа, 2024. — 469 с. — Режим доступа https://www.iprbookshop.ru/133907.

3. Пальмов С.В. Интеллектуальный анализ данных [Электронный ресурс]: учеб. пособие. – Самара: Поволжский государственный университет телекоммуникаций и информатики, 2017. – 127 с. – Режим доступа http://www.iprbookshop.ru/75376

4. Протодьяконов, А. В. Алгоритмы Data Science и их практическая реализация на Python [Электронный ресурс]: учебное пособие / А. В. Протодьяконов, П. А. Пылов, В. Е. Садовников. — Москва, Вологда : Инфра-Инженерия, 2022. — 392 с. — Режим доступа: https://www.iprbookshop.ru/124000

5 Лэм, Ч. Hadoop в действии [Электронный ресурс]/ Ч. Лэм. — Москва : ДМК Пресс, 2019. — 424 с. —Режим доступа https://www.iprbookshop.ru/124537

8.2. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень лицензионного программного обеспечения, ресурсов информационно-телекоммуникационной сети «Интернет», современных профессиональных баз данных и информационных справочных систем:

Программное обеспечение

Hadoop (Hortonworks Data platform)
Apache Spark
Apache Kafka
Apache Hive

MongoDB
HBase
Redis
IntellijIdea community (java)
OpenJDK11
Python

Ресурсы информационно-телекоммуникационной сети «Интернет»
http://www.edu.ru/
Образовательный портал ВГТУ

Информационные справочные системы
http://window.edu.ru
https://wiki.cchgeu.ru/

Современные профессиональные базы данных
https://habr.com/ru/
https://sources.ru/
https://proglib.io/

9 МАТЕРИАЛЬНО-ТЕХНИЧЕСКАЯ БАЗА, НЕОБХОДИМАЯ ДЛЯ ОСУЩЕСТВЛЕНИЯ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА

Для проведения обучения по дисциплине используются:
Компьютерный класс
Учебная аудитория для проведения учебных занятий, включаюшая:
– рабочее место преподавателя (стол, стул);
– рабочие места обучающихся (столы, стулья)
– персональные компьютеры с установленным ПО, подключенные к сети Интернет (12 шт.);
– принтер;
– доска магнитно-маркерная поворотная
- оборудование для лекционных демонстраций и проекционная аппаратура.

Помещение для самостоятельной работы. Читальный зал с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду.

10. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

По дисциплине «Технологии обработки больших данных» читаются

лекции, проводятся лабораторные работы, выполняется курсовой проект.

Основой изучения дисциплины являются лекции, на которых излагаются наиболее существенные и трудные вопросы, а также вопросы, не нашедшие отражения в учебной литературе.

Лабораторные работы выполняются на лабораторном оборудовании в соответствии с методиками, приведенными в указаниях к выполнению работ.

Методика выполнения курсового проекта изложена в учебно-методическом пособии. Выполнять этапы курсового проекта должны своевременно и в установленные сроки.

Контроль усвоения материала дисциплины производится проверкой курсового проекта, защитой курсового проекта.

Вид учебных занятий	Деятельность студента
Лекция	Написание конспекта лекций: кратко, схематично, последовательно фиксировать основные положения, выводы, формулировки, обобщения; помечать важные мысли, выделять ключевые слова, термины. Проверка терминов, понятий с помощью энциклопедий, словарей, справочников с выписыванием толкований в тетрадь. Обозначение вопросов, терминов, материала, которые вызывают трудности, поиск ответов в рекомендуемой литературе. Если самостоятельно не удается разобраться в материале, необходимо сформулировать вопрос и задать преподавателю на лекции или на практическом занятии.
Лабораторная работа	Лабораторные работы позволяют научиться применять теоретические знания, полученные на лекции при решении конкретных задач. Чтобы наиболее рационально и полно использовать все возможности лабораторных для подготовки к ним необходимо: следует разобрать лекцию по соответствующей теме, ознакомится с соответствующим разделом учебника, проработать дополнительную литературу и источники, решить задачи и выполнить другие письменные задания.
Самостоятельная работа	Самостоятельная работа студентов способствует глубокому усвоения учебного материала и развитию навыков самообразования. Самостоятельная работа предполагает следующие составляющие: - работа с текстами: учебниками, справочниками, дополнительной литературой, а также проработка конспектов лекций; - выполнение домашних заданий и расчетов; - работа над темами для самостоятельного изучения; - участие в работе студенческих научных конференций, олимпиад; - подготовка к промежуточной аттестации.
Подготовка к промежуточной аттестации	Готовиться к промежуточной аттестации следует систематически, в течение всего семестра. Интенсивная подготовка должна начаться не позднее, чем за месяц-полтора до промежуточной аттестации. Данные перед экзаменом, экзаменом три дня эффективнее всего использовать для повторения и систематизации материала.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Перечень вносимых изменений	Дата внесения изменений	Подпись заведующего кафедрой, ответственной за реализацию ОПОП