

Федеральное государственное образовательное бюджетное учреждение  
высшего образования  
**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ  
РОССИЙСКОЙ ФЕДЕРАЦИИ»**  
(Финансовый университет)

**Краснодарский филиал Финуниверситета**

Кафедра «Математика и информатика»

СОГЛАСОВАНО

ООО «Портал-Юг»  
Генеральный директор



Е.В. Мостовой

«20» февраля 2024 г.

УТВЕРЖДАЮ

Краснодарский филиал  
Финуниверситета

Директор



Э.В.Соболев

«20» февраля 2024 г.

Калайдин Е.Н.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ  
ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ**  
студентов, обучающихся по направлению подготовки  
01.03.02 Прикладная математика и информатика  
в соответствии с образовательными стандартами Финансового университета  
(программа подготовки бакалавров)

*Рекомендовано Ученым советом Краснодарского филиала Финуниверситета  
(протокол № 12 от 20.02.2024)*

*Одобрено кафедрой «Математика и информатика»  
(протокол № 13 от 27.02.2024)*

**Краснодар 2024**

УДК: 004.9  
ББК: 32.97  
К17

Рецензенты: Е.А. Демехин, профессор кафедры «Математика и информатика» Краснодарского филиала Финансового университета при Правительстве РФ. В.А. Кирий кандидат физико-математических наук, доцент кафедры «Математика и информатика» Краснодарского филиала Финуниверситета.

Калайдин Е.Н. Рабочая программа дисциплины технологии обработки больших данных для обучающихся по направлению 01.03.02 Прикладная математика и информатика, профиль «Анализ данных и принятие решений в экономике и финансах». – Краснодар: Краснодарский филиал Финуниверситета, кафедра «Математика и информатика», 2024 г.

Дисциплина Технологии обработки больших данных относится к предпрофильному профессиональному циклу по направлению подготовки 01.03.02-Прикладная математика и информатика.

В рабочей программе дисциплины определены ее цель, требования к результатам освоения дисциплины, содержание программы, тематика аудиторных занятий, формы самостоятельной работы, оценочные средства для текущего контроля и промежуточной аттестации, учебно-методическое и информационное обеспечение.

Рабочая программа дисциплины технологии обработки больших данных

*Формат 60\*90/16. Гарнитура Times New Roman*

*Усл. п.л. 2,0. Изд. № \_от.*

*Тираж 100 экз.*

*Заказ № .*

*Отпечатано в Краснодарском филиале Финуниверситета*

© Калайдин Е.Н.  
© Краснодарский филиал Финуниверситета, 2024

## Содержание

1.Наименование дисциплины .....	4
2.Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине .....	4
3.Место дисциплины в структуре образовательной программы .....	5
4.Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся .....	5
5.Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий .....	6
5.1.Содержание дисциплины .....	6
5.2.Учебно-тематический план .....	9
5.3.Содержание семинаров, практических занятий .....	13
6.Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине .....	15
6.1.Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы .....	15
6.2.Перечень вопросов, заданий, тем для подготовки к текущему контролю .....	17
7.Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине .....	19
7.1.Типовые контрольные задания или иные материалы, необходимые для оценки индикаторов достижения компетенций, умений и знаний .....	20
7.2.Примерные вопросы для подготовки к экзамену .....	23
7.3.Пример экзаменационного билета .....	36
8.Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины .....	38
9.Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины .....	39
10.Методические указания для обучающихся по освоению дисциплины .....	39
11.Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем .....	40
12.Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине .....	40

## 1. Наименование дисциплины

Дисциплина «Технологии обработки больших данных»

## 2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине

Дисциплина «Технологии обработки больших данных» обеспечивает инструментарий формирования следующих компетенций: ПКН-4.

Код компетенции	Наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
ПКН-4	Способен проектировать и реализовывать прикладные программные системы в соответствии с анализом задачи и требований к ним	1. Демонстрирует базовые знания о существующих математических методах и системах программирования.	<b>Знать:</b> существующие стандарты, необходимые для создания технического задания и технического проекта с учетом специфических требований больших данных <b>Уметь:</b> использовать и адаптировать существующие стандарты с учетом специфических требований больших данных
		2. Использует и адаптирует существующие математические методы и системы программирования для решения прикладных задач.	<b>Знать:</b> технологию разработки технических заданий и технических проектов, в которых используются технологии больших данных <b>Уметь:</b> разрабатывать технические задания и технических проекты, в которых используются технологии больших данных
		3. Владеет навыками проектирования и разработки компонентов программного обеспечения на основе современных парадигм, технологий и языков программирования.	<b>Знать:</b> современные принципы управления рабочими проектами, применяемыми к технологической инфраструктуре больших данных <b>Уметь:</b> применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных

		4. Применяет методы машинного обучения для решения прикладных задач анализа данных.	<b>Знать:</b> методы и инструменты анализа данных и машинного обучения <b>Уметь:</b> применять методы и инструменты анализа данных
--	--	---	---

### 3. Место дисциплины в структуре образовательной программы

Дисциплина «Технологии обработки больших данных» является предпрофильным профессиональным циклом профиля «Анализ данных и принятие решений в экономике и финансах» по направлению подготовки 01.03.02 «Прикладная математика и информатика».

### 4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся

#### Очная форма обучения

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 5 (в часах)
Общая трудоемкость дисциплины	5/144	144
Контактная работа - Аудиторные занятия	50	50
Лекции	16	16
Семинарские, практические занятия	34	34
Самостоятельная работа	94	94
Вид текущего контроля	Контрольная работа	
Вид промежуточной аттестации	Экзамен	

#### Очно – заочная форма обучения

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 5 (в часах)
Общая трудоемкость дисциплины	5/180	180
Контактная работа - Аудиторные занятия	28	28
Лекции	12	12
Семинарские, практические занятия	16	16
Самостоятельная работа	152	152
Вид текущего контроля	Контрольная работа	
Вид промежуточной аттестации	Экзамен	

## **5.Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий**

### **5.1.Содержание дисциплины**

#### **Тема 1. Библиотека NumPy и Pandas.**

В рамках темы рассматривается технологический стек Python для обработки и анализа данных, возможности Python как glue language, специфика библиотеки NumPy и ее роль в экосистеме Python. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с единичными исходными данными. Универсальные функции и применение функций по осям в NumPy. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры Маскирование и прихотливое индексирование в NumPy.

В рамках темы рассматриваются возможности библиотеки Pandas. Организация Pandas DataFrame и организация индексации для DataFrame и Series; применение универсальных функций и работа с пустыми значениями в Pandas. Объединение данных из нескольких Pandas DataFrame: общая логика и примеры. Рассматривается операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение».

#### **Тема 2. Использование различных форматов файлов в задачах обработки данных.**

В рамках темы рассматриваются принципы работы с файлами, файлы и операционные системы. Специфика текстовых и бинарных файлов.

В рамках темы рассматривается задача сериализации и десериализации данных и использование различных форматов файлов для ее решения. Описание формата файла JSON и пример описания данных в этом формате и взаимодействия с ним в Python.

В рамках темы рассматриваются формат XML и модель DOM: общая характеристика, пример описания данных в XML и DOM, работа с ними с помощью библиотеки BeautifulSoup.

В рамках темы рассматривается проблематика форматов файлов для хранения и обработки больших данных. Форматы файлов NPY и HDF: общая характеристика, пример взаимодействия с данными этих форматов в Python.

#### **Тема 3. Взаимодействие с табличными данными в приложениях обработки данных.**

В рамках темы рассматривается формат файлов CSV, представление данных в этом формате и взаимодействие с ним в Python.

В рамках темы рассматриваются возможности использования Excel для внешних приложений обработки данных. Взаимодействие с Excel из Python с помощью библиотеки XLWings: принципы работы и примеры использования.

#### **Тема 4. Визуализация данных.**

В рамках темы рассматриваются основы работы с библиотекой `matplotlib`: организация системы координат, оформление осей, цвета и цветовые карты в `matplotlib`, стили линий и маркеры. `Pyplot` и объектно-ориентированный интерфейс `matplotlib`. Управление фигурами и создание множества графиков на одном рисунке. Различные типы графиков.

В рамках темы рассматривается визуализация данных с помощью библиотеки `Pandas`: набор методов для построения графиков, реализованный в структурах `Series` и `DataFrame`.

В рамках темы проводится введение в разведочный анализ данных: типы признаков, анализ распределений, анализ мер центральной тенденции и поиск выбросов, анализ взаимного распределения и парных корреляций. Проведение разведочного анализа данных с помощью библиотеки `Seaborn`.

#### **Тема 5. Работа со строками в приложениях обработки данных.**

В рамках темы рассматриваются возможности `python` по форматированию строк: %-форматирование, метод `format`, f-строки.

В рамках темы рассматриваются основы работы с регулярными выражениями: базовый синтаксис, примеры. Модуль `re` в `Python`. Примеры использования регулярных выражений.

В рамках темы рассматривается использования хэширования при работе со строками. Строки в библиотеке `numpy`.

#### **Тема 6. Введение в обработку текста на естественном языке в задачах обработки данных.**

В рамках темы рассматриваются сегментация и токенизация текста на естественном языке, стемминг и лемматизация, примеры на `Python`. Использование мемоизации на примере работы со строками. Расстояние Ле-венштейна: определение, алгоритм эффективного поиска оптимального редакционного предписания, пример поиска на `Python`. Векторное представление текста на естественном языке: общий алгоритм подходов `TF`; `TF-IDF`.

#### **Тема 7. Профилирование процессов обработки данных, библиотека `Numba` и векторизация в `Numpy` и `Numba`.**

В рамках темы рассматривается профилирование реализации алгоритмов на `Python`, принципы решения задачи оптимизации производительности алгоритма. Библиотека `Numba`: принципы работы, базовые примеры использования. Векторизация в `numpy`: ключевые параметры функции, примеры применения, использование обобщенной сигнатуры функции.

#### **Тема 8. Взаимодействие с базой данных в приложениях обработки данных.**

В рамках темы рассматривается взаимодействие из `Python` с базой данных на примере `API SQLite`. Базовые возможности работы с транзакциями.

#### **Тема 9. Параллельная обработка данных.**

В рамках темы рассматривается специфика современного аппаратного

обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение.

Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. Специфика различия между потоками и процессами.

Проблема Global Interpreter Lock в Python и способы обхода ее ограничений. Модуль Python multiprocessing – назначение и основные возможности, API multiprocessing.Pool.

## **Тема 10. Библиотека Dask**

В рамках темы рассматривается библиотека для анализа больших объемов данных Python Dask, различные предлагаемые ей подходы к обработке данных. В частности, три ключевых структуры данных Dask: Dask.Array, Dask.DataFrame и Dask.Bag их специфика и принцип выбора структур данных при решении задач. Рассматривается граф зависимостей задач, как ключевая структура для организации параллельной обработки данных в Python Dask. Рассматривается принцип и примеры использования распараллеливание алгоритмов с помощью dask.delayed .

Рассматривается структура данных Dask.Array, специфика ее реализации и применения, процедура создания, поддерживаемые Dask.Array операции и ее отличия от NumPy ndarray. Рассматривается структура данных Dask.DataFrame, специфика ее реализации и применения, процедура создания, ограничения использования Dask.DataFrame. Рассматриваются операции мэппинга в Dask.DataFrame и операции Dask.DataFrame работающие со скользящим окном. Рассматривается структура данных Dask.Bag, специфика ее реализации и применения, процедура создания, поддерживаемые Dask.Bag операции. Организация вычислений с помощью Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных с помощью Dask.Bag.

## **Тема 11. Обзор проблем обработки больших данных и вычисления общего назначения на GPU**

Большие данные – определение и причины возникновения задач обработки больших данных. Вызовы «Больших данных»: объем данных, слабая структурированность данных, связность данных, обработка данных с помощью независимых сервисов. Специфика аппаратного обеспечения для решения задач обработки больших данных. Проблема выбора типичных средств обработки данных, адекватных различным объемам данных. Принцип обработки данных на базе операций map / filter / reduce, принципы архитектуры hadoop. Источники больших данных и прикладные задачи обработки больших данных.

История развития и общая характеристика GPU. Архитектура Nvidia CUDA. Принципы организации вычислений в архитектуре Nvidia CUDA.

Знакомство с библиотекой PyTorch. Понятие тензора в PyTorch. Базовые операции с тензорами в PyTorch.



## 5.2. Учебно-тематический план

### Очная форма обучения

№ п/ п	Наименование темы (раздела) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемости и
		Всего	Контактная работа Аудиторная работа			Самост оатель ная работа	
			Обща я, в т. ч.:	Ле кц ии	Семин а-ры, практи ческие заняти я		
1	Библиотека NumPy и Pandas	13	5	2	3	8	Участие в решении задач на практически х занятиях. Обсуждения по результатам самостоятель ной работы
2	Использован ие различных форматов файлов в задачах обработки данных.	15	5	2	3	10	
3	Взаимодейст вие с табличными данными в приложениях обработки данных.	12	4	1	3	8	
4	Визуализаци я данных	12	4	1	3	8	
5	Работа со строками в приложениях обработки данных	12	4	1	3	8	
6	Введение в обработку текста на естественно м языке в задачах обработки данных	13	4	1	3	9	

7	Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba.	12	4	1	3	8	
8	Взаимодействие с базой данных в приложениях обработки данных.	12	4	1	3	8	
9	Параллельная обработка данных	15	5	1	4	10	
10	Библиотека Dask	16	7	4	3	9	
11	Обзор проблем обработки больших данных и вычисления общего назначения на GPU	12	4	1	3	8	
В целом по дисциплине		144	68	16	34	94	Согласно учебному плану: контрольная работа

Очно – заочная форма обучения

№ п/ п	Наимено вание темы (раздела) дисципли ны	Трудоемкость в часах					Формы текущего контроля успеваемост и
		Всего	Контактная работа - Аудиторная работа			Самост оатель ная работа	
			Общ ая, в т. ч.:	Ле кц ии	Семин а-ры, практи ческие заняти я		

1	Библиотека NumPy и Pandas	13	3	1	2	10	Участие в решении задач на практических занятиях. Обсуждения по результатам самостоятельной работы
2	Использование различных форматов файлов в задачах обработки данных.	13	2	1	1	11	
3	Взаимодействие с табличными данными в приложениях обработки данных.	13	2	1	1	11	
4	Визуализация данных	12	2	1	1	10	
5	Работа со строками в приложениях обработки данных	12	3	1	2	11	
6	Введение в обработку текста на естественном языке в задачах обработки данных	12	2	1	1	10	
7	Профилирование процессов обработки данных, библиотека Numba и векторизация в NumPy и Numba.	13	2	1	1	11	

8	Взаимодействие с базой данных в приложениях обработки данных.	14	3	1	2	11	
9	Параллельная обработка данных	13	3	1	2	11	
10	Библиотека Dask	13	3	2	1	10	
11	Обзор проблем обработки больших данных и вычисления общего назначения на GPU	13	3	1	2	10	
В целом по дисциплине		180	28	12	16	152	Согласно учебному плану: контрольная работа

### 5.3.Содержание семинаров, практических занятий

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарских, практических занятиях, рекомендуемые источники из разделов 8, 9 (указывается раздел и порядковый номер источника)	Формы проведения занятий
Библиотека NumPy и Pandas	Технологический стек Python для обработки и анализа данных Возможности Python как glue language Организация массивов в NumPy: хранение данных, создание массивов Принципы реализации операций с едиными исходными данными. Универсальные функции и применение функций по осям в NumPy. Организация Pandas DataFrame и организация индексации для DataFrame и Series. Применение универсальных функций и работа с пустыми значениями в Pandas. Объединение данных из нескольких Pandas DataFrame: общая логика и примеры. 8[1], 10[1], 9[2]	Интерактивная форма, работа на компьютере
Использование различных форматов файлов в задачах обработки данных	Формат файлов Pickle, представление данных в этом формате и взаимодействие с ним в Python. Формат файлов JSON, представление данных в этом формате и взаимодействие с ним в Python. Формат XML и модель DOM: общая характеристика, пример описания данных в XML и DOM Работа с XML с помощью библиотеки BeautifulSoup. 8[1], 9[1], 9[2]	Интерактивная форма, работа на компьютере Интерактивная форма, работа на компьютере
Взаимодействие с табличными данными в приложениях обработки данных.	Взаимодействие с Excel из Python с помощью библиотеки XLWings. Формат файлов CSV, представление данных в этом формате и взаимодействие с ним в Python 8[1], 9[1], 9[2]	Интерактивная форма, работа на компьютере
Визуализация данных	Построение визуализаций с помощью библиотеки matplotlib Построение визуализаций с помощью библиотеки pandas Построение визуализаций с помощью библиотеки seaborn 8[1], 9[1], 9[2]	Интерактивная форма, работа на компьютере
Работа со строками в приложениях обработки данных	Основы работы с регулярными выражениями: базовый синтаксис, примеры. Модуль re в Python. 8[1], 9[1],	Интерактивная форма, работа на компьютере

	9[2]	
Введение в обработку текста на естественном языке в задачах обработки данных.	Сегментация и токенизация текста на естественном языке, стемминг и лемматизация, примеры на Python. Расстояние Левенштейна: определение, алгоритм эффективного поиска оптимального редакционного предписания, пример поиска на Python. 9[1], 8[1], 8[2]	Интерактивная форма, работа на компьютере
Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba	профилирование реализации алгоритмов на Python принципы решения задачи оптимизации производительности алгоритма Библиотека Numba: принципы работы, базовые примеры использования. 9[1], 8[1], 8[2]	Интерактивная форма, работа на компьютере
Взаимодействие с базой данных приложениях обработки данных	Взаимодействие из Python с базой данных с помощью API SQLite. 9[1], 8[1], 8[2]	Интерактивная форма, работа на компьютере
Параллельная обработка данных	специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения. Проблема Global Interpreter Lock в Python и способы обхода ее ограничений. Модуль Python multiprocessing – назначение и основные возможности, API multiprocessing.Pool. 9[1], 8[1], 8[2]	Интерактивная форма, работа на компьютере
Библиотека Dask	Подход к обработке данных с помощью библиотеки Dask. Структура данных Dask.Array – принцип работы, API, примеры использования. Структура данных Dask.DataFrame – принцип работы, API, примеры использования. Структура данных Dask.Bag – принцип работы, API, примеры использования. 9[1], 8[1], 8[2]	Интерактивная форма, работа на компьютере

Обзор проблем обработки больших данных и вычисления общего назначения на GPU	Вызовы «Больших данных»: объем данных, слабая структурированность данных, связность данных, обработка данных с помощью независимых сервисов. Источники больших данных и прикладные задачи обработки больших данных. Архитектура Nvidia CUDA. Принципы организации вычислений в архитектуре Nvidia CUDA. 9[1], 8[1], 8[2]	Интерактивная форма, работа на компьютере
--	---	---

## 6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

### 6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы

Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы
Библиотека NumPy и Pandas	Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры. Маскирование и прихотливое индексирование в NumPy. Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение».	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Использование различных форматов файлов в задачах обработки данных	Формат файлов NPY, представление данных в этом формате и взаимодействие с ним в Python. Формат файлов HDF, представление данных в этом формате и взаимодействие с ним в Python.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Взаимодействие с табличными данными в приложениях обработки данных.	Продвинутое взаимодействие с Excel из Python с помощью библиотеки XLWings.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.

Визуализация данных	Построение трехмерных графиков Продвинутая работа с цветовыми картами	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Работа со строками в приложениях обработки данных	Использования хэширования при работе со строками. Строки в библиотеке numpy.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Введение в обработку текста на естественном языке в задачах обработки данных.	Использование мемоизации на примере работы со строками. Векторное представление текста на естественном языке: общий алгоритм подходов TF; TF-IDF.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba	Векторизация в numpy: ключевые параметры функции, примеры применения Использование обобщенной сигнатуры функции в numpy и numba.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Взаимодействие с базой данных в приложениях обработки данных	Базовые возможности работы с транзакциями с помощью API SQLite.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Параллельная обработка данных	Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. Специфика различия между потоками и процессами. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.



Библиотека Dask	Организация вычислений с помощью Map / Filter / Reduce: общий принцип и специфика параллельной реализации обработки данных с помощью Dask.Bag. Организация вычислений с помощью API Dask Delayed.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Обзор проблем обработки больших данных и вычисления общего назначения на GPU	Специфика аппаратного обеспечения для решения задач обработки больших данных. Проблема выбора типичных средств обработки данных, адекватных различным объемам данных. Знакомство с библиотекой PyTorch. Понятие тензора в PyTorch. Базовые операции с тензорами в PyTorch.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.

## 6.2.Перечень вопросов, заданий, тем для подготовки к текущему контролю

### Примерные вопросы к контрольной работе

1. Большие данные – определение и причины возникновения задач обработки больших данных
2. Специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений
3. Выбор типичных средств обработки данных, адекватных различным объемам данных; принцип обработки данных на базе операций map / filter / reduce
4. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение
5. Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения
6. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем
7. Различия между потоками и процессами, различие между различными планировщиками в Dask
8. Граф зависимостей задач – суть структуры данных, ее построение и использование в Dask
9. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач
10. Dask.Array – структура данных, специфика реализации и применения, процедура создания
11. Dask.Array – поддерживаемые операции и отличия от NumPy ndarray
12. Распараллеливание алгоритмов с помощью dask.delayed – принцип и примеры использования

13. Дополнительные параметры декоратора `dask.delayed` – назначение и примеры использования
14. Использование `dask.delayed` для объектов и операции над объектами `dask.delayed`, включая ограничения их использования
15. `Dask.DataFrame` - структура данных, специфика реализации и применения, процедура создания `Dask.DataFrame`
16. Ограничения использования `Dask.DataFrame` и операции мэппинга в `Dask.DataFrame`
17. Поддержка `Dask.DataFrame` операций, работающих со скользящим окном
18. Совместное использование промежуточных результатов в Dask: принцип работы и примеры использования
19. `Dask.Bag` - структура данных, специфика реализации и применения, процедура создания `Dask.Bag`
20. Организация вычислений с помощью Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных в `Dask.Bag`
21. API `Dask.Bag` – функции мэппинга, фильтрации и преобразования

## **Примерные задания контрольной работы**

### **Задание 1**

1. В массиве чисел, хранящихся в файле `finance.hdf5`, найти строку (вывести ее индекс и содержащиеся значения), в которой более всего значений, превышающих среднее значение по всему массиву. Для расчётов использовать `dask.array`.
2. В массиве чисел, хранящихся в файле `finance.hdf5`, подсчитать количество строк, в которых более 600 значений больше среднего значения по всему массиву. Для расчётов использовать `dask.array`.
3. В массиве чисел, хранящихся в файле `finance.hdf5`, подсчитать количество значений, не отклоняющихся от среднего значения более чем на 3 стандартных отклонения. Для расчетов использовать `dask.array`

### **Задание 2**

1. В `accounts*.csv` найти `id`, для которого в столбце `amount` встречается наибольшее количество значений, кратных трем. Выполнить задание с использованием Dask, распараллелив процесс обработки данных
2. В `accounts*.csv` найти `id`, для которого сумма положительных значений в столбце `amount` наибольшая. Выполнить задание с использованием Dask, распараллелив процесс обработки данных.
3. В `accounts*.csv` найти `id`, для которого в столбце `amount` встречается наибольшее количество значений между 1000 и 1500. Выполнить задание с использованием Dask, распараллелив процесс обработки данных.

### **Задание 3**

Датасет: `all_k.zip`

Подсчитать, сколько раз в текстовых файлах, лежащих в `all_k.zip`, встречаются предложения трех видов: вопросительные (в окончании имеют вопросительный знак), побудительные (в окончании имеют восклицательный знак и не имеют вопросительного) и повествовательные (в окончании имеют точку или троеточие, при

этом нужно исключить учет точек, встречающихся в сокращениях, таких как "т.к.").

Выполнить задание с использованием Dask (корректным!), распараллелив процесс обработки данных (использование Dask должно приводить к истинной параллельной обработке данных).

#### **Задание 4**

Датасет: all\_k.zip

Подсчитать, сколько раз встречается каждое из личных местоимений в именительном падеже (полный список: я, ты, он, она, оно, мы, вы, они) в текстовых файлах, лежащих в папке: all\_k.zip.

Выполнить задание с корректным использованием Dask, распараллелив процесс обработки данных (использование Dask должно приводить к истинной параллельной обработке данных).

### **Примерная тематика курсового проекта**

1. Прогнозная аналитика и моделирование объемов продаж акций
2. Визуализация аналитических данных в области макроэкономики
3. Визуализация аналитических данных Московской биржи
4. Использование технологии больших данных для анализа портфельных рисков
5. Использование параллельных вычислений реализации численных методов решения математических задач
6. Анализ и сравнение различных фреймворков для визуализации данных
7. Применение распределенных вычислений и экосистемы Hadoop для решения задачи анализа данных
8. Анализ больших данных для построения прогнозов на рынке ценных бумаг
9. Использование больших данных для оценки кредитоспособности контрагентов на основе анализа текстов новостей
10. Проведение анализа собранных из внешних источников данных

### **Критерии балльной оценки различных форм текущего контроля успеваемости**

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях кафедры «Математика и информатика».

### **7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине**

Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы содержится в разделе 1. «Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине».

## 7.1. Типовые контрольные задания или иные материалы, необходимые для оценки индикаторов достижения компетенций, умений и знаний

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»	
<b>ПКН-4 Способен проектировать и реализовывать прикладные программные системы в соответствии с анализом задачи и требований к ним</b>					
Демонстрирует базовые знания о существующих математических методах и системах программирования					
Знать: существующие стандарты, необходимые для создания технического задания и технического проекта с учетом специфических требований больших данных	Фрагментарное представление о существующих стандартах, необходимы х для создания технического задания и технического проекта с учетом специфических требований больших данных	Неполные представления о существующих стандартах, необходимы для создания технического задания и технического проекта с учетом специфических требований больших данных	Сформированные, но содержащие отдельные пробелы представления о существующих стандартах, необходимы х для создания технического задания и технического проекта с учетом специфических требований больших данных	Сформированные систематические представления о существующих стандартах, необходимы х для создания технического задания и технического проекта с учетом специфических требований больших данных	Вопросы для оценки знаний и умений, задания в виде расчетных задач, тестовые задания
Уметь: использовать и адаптировать существующие стандарты с учетом специфических требований больших данных	Фрагментарное умение использовать и адаптировать существующие стандарты с учетом специфических требований больших данных	Несистематическое применение умений использовать и адаптировать существующие стандарты с учетом специфических требований больших данных	В целом успешное, но содержащее отдельные пробелы умение использовать и адаптировать существующие стандарты с учетом специфических	Сформированное умение использовать и адаптировать существующие стандарты с учетом специфических требований больших данных	Вопросы для оценки знаний и умений,

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетвор ительно»	«удовлетвор ительно»	«хорошо»	«отлично»	
		данных	их требований больших данных		
Использует и адаптирует существующие математические методы и системы программирования для решения прикладных задач					
Знать: технологии разработки технических заданий и технических проектов, в которых используютс я технологии больших данных	Фрагментар ное представл ение о технологии разработки технических заданий и технических проектов, в которых используютс я технологии больших данных	Неполные представлен ия о технологии разработки технических заданий и технических проектов, в которых используютс я технологии больших данных	Сформирова нные, но содержащие отдельные пробелы представлен ия о технологии разработки технических заданий и технических проектов, в которых используютс я технологии больших данных	Сформирова нные систематиче ские представлен ия о технологии разработки технических заданий и технических проектов, в которых используютс я технологии больших данных	Вопросы для оценки знаний и умений, задания в виде расчетных задач, тестовые задания
Уметь: разрабатыва ть технические задания и технических проекты, в которых используютс я технологии больших данных	Фрагментар ное умение использоват ь и адаптироват ь существующ ие стандарты с учетом специфическ их требований больших данных	Несистемати ческое применение умений использоват ь и адаптироват ь существующ ие стандарты с учетом специфическ их требований больших данных	В целом успешное, но содержащее отдельные пробелы умение использоват ь и адаптироват ь существующ ие стандарты с учетом специфическ их требований больших данных	Сформирова нное умение использоват ь и адаптироват ь существующ ие стандарты с учетом специфическ их требований больших данных	Задания в виде расчетных задач, тестовые задания
Владеет навыками проектирования и разработки компонентов программного обеспечения на основе современных парадигм, технологий и языков программирования					

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»	
Знать: современные принципы управления рабочими проектами, применяемыми к технологической инфраструктуре больших данных	Фрагментарное представление о современных принципах управления рабочими проектами, применяемых к технологической инфраструктуре больших данных	Неполное представление о современных принципах управления рабочими проектами, применяемых к технологической инфраструктуре больших данных	Сформированные, но содержащие отдельные пробелы представления о современных принципах управления рабочими проектами, применяемых к технологической инфраструктуре больших данных	Сформированные систематические представления о современных принципах управления рабочими проектами, применяемых к технологической инфраструктуре больших данных	Вопросы для оценки знаний и умений, задания в виде расчетных задач, тестовые задания
Уметь: применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных	Фрагментарное умение применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных	Несистематическое применение современных принципов управления рабочими проектами технологической инфраструктуры больших данных	В целом успешное, но содержащее отдельные пробелы умение применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных	Сформированное умение применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных	Задания в виде расчетных задач, тестовые задания
Применяет методы машинного обучения для решения прикладных задач анализа данных					
Знать: методы и инструменты анализа данных и машинного	Фрагментарное представление о методах и инструментах	Неполное представление о методах и инструментах анализа	Сформированные, но содержащие отдельные пробелы представления	Сформированные систематические представления о методах	Вопросы для оценки знаний и умений, задания в виде расчетных

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»	
обучения	х анализа данных и машинного обучения	данных и машинного обучения	ия о методах и инструментах анализа данных и машинного обучения	и инструментах анализа данных и машинного обучения	задач, тестовые задания
Уметь: применять методы и инструменты анализа данных	Фрагментарное умение применять методы и инструменты анализа данных	Несистематическое применение методов и инструментов в анализа данных	В целом успешное, но содержащее отдельные пробелы умение применять методы и инструменты анализа данных	Сформированное умение применять методы и инструменты анализа данных	Задания в виде расчетных задач, тестовые задания

## 7.2. Вопросы для оценки знаний и умений, характеризующих формирование компетенций

Шифр компетенции	Вопросы	Правильный ответ
ПKN - 4	1. Дайте определение большим данным и объясните семь Против больших данных.	Большие данные - это набор больших и сложных полуструктурированных и неструктурированных наборов данных, которые потенциально могут предоставлять полезную информацию с использованием традиционных инструментов управления данными. Операции с большими данными требуют специализированных инструментов и методов, поскольку реляционная база данных не может управлять таким большим объемом данных. Большие данные позволяют предприятиям глубже понять свою отрасль и помогают им

		<p>извлекать ценную информацию из регулярно собираемых неструктурированных и необработанных данных. Большие данные также позволяют предприятиям принимать более обоснованные бизнес-решения.</p> <p>Семь Против больших данных - это</p> <p>Объем: Объем представляет собой объем данных, растущий экспоненциально. Пример: петабайты и эксабайты.</p> <p>Скорость: Скорость представляет собой скорость, с которой растут данные.</p> <p>Разнообразие: Разнообразие относится к типам данных в различных форматах данных, включая текст, аудио и видео.</p> <p>Ценность: Ценность означает получение ценной информации для удовлетворения потребностей бизнеса и получения доходов.</p> <p>Правдивость: Правдивость связана с точностью анализируемых данных. Это относится к тому, насколько надежны данные или, другими словами, к качеству анализируемых данных.</p> <p>Визуализация: Визуализация относится к представлению данных руководству для целей принятия решений.</p> <p>Изменчивость: изменчивость относится к данным, которые постоянно меняются.</p>
	<p>2. Как развернуть модель больших данных? Укажите ключевые шаги.</p>	<p>Развертывание модели больших данных включает в себя три этапа:</p> <p>Прием данных: Это первый шаг в развертывании модели больших данных - прием данных, то есть извлечение данных из нескольких источников. Этот процесс включает в себя сбор данных из нескольких источников, таких как сайты социальных сетей,</p>



		<p>корпоративное программное обеспечение и файлы журналов.</p> <p>Хранение данных: Следующим шагом после приема данных является их хранение в HDFS или базе данных NoSQL, такой как HBase. Хранилище HBase идеально подходит для случайных операций чтения / записи, в то время как HDFS предназначена для последовательных процессов.</p> <p>Обработка данных: Это заключительный шаг в развертывании модели больших данных. Как правило, обработка данных выполняется с использованием таких фреймворков, как Hadoop, Spark, MapReduce, Flink и Pig, и это лишь некоторые из них.</p>														
	<b>3. Как Hadoop связан с большими данными?</b>	Hadoop - это платформа с открытым исходным кодом для хранения, анализа и интерпретации больших объемов неструктурированных данных с целью получения ценной информации для принятия более эффективных бизнес-решений.														
	4. Объясните разницу между Hadoop и RDBMS.	<table><tr><th>Ключевые особенности</th><th>Hadoop</th></tr><tr><td>Обзор</td><td>Hadoop - это программное обеспечение с открытым исходным кодом, которое объединяет тысячи компьютеров, требующих больших объемов данных и их обработки.</td></tr><tr><td>Разнообразие данных</td><td>Hadoop хранит структурированные, полуструктурированные и неструктурированные данные.</td></tr><tr><td>Хранение данных</td><td>Hadoop хранит данные.</td></tr><tr><td>Аппаратное обеспечение</td><td>Hadoop использует оборудование.</td></tr><tr><td>Масштабируемость</td><td>Hadoop обладает масштабируемостью.</td></tr><tr><td>Пропускная способность</td><td>Высокий</td></tr></table>	Ключевые особенности	Hadoop	Обзор	Hadoop - это программное обеспечение с открытым исходным кодом, которое объединяет тысячи компьютеров, требующих больших объемов данных и их обработки.	Разнообразие данных	Hadoop хранит структурированные, полуструктурированные и неструктурированные данные.	Хранение данных	Hadoop хранит данные.	Аппаратное обеспечение	Hadoop использует оборудование.	Масштабируемость	Hadoop обладает масштабируемостью.	Пропускная способность	Высокий
Ключевые особенности	Hadoop															
Обзор	Hadoop - это программное обеспечение с открытым исходным кодом, которое объединяет тысячи компьютеров, требующих больших объемов данных и их обработки.															
Разнообразие данных	Hadoop хранит структурированные, полуструктурированные и неструктурированные данные.															
Хранение данных	Hadoop хранит данные.															
Аппаратное обеспечение	Hadoop использует оборудование.															
Масштабируемость	Hadoop обладает масштабируемостью.															
Пропускная способность	Высокий															

5. Назовите несколько методов обработки больших данных.	Методы обработки больших данных включают Обработка потоков больших данных Пакетная обработка больших данных Обработка больших данных в режиме реального времени
6. Что такое "выброс" в контексте больших данных?	Выбросы - это точки данных, которые очень удалены от группы и не принадлежат ни к каким кластерам или группам. Наличие выбросов влияет на поведение модели; они предсказывают неправильные результаты или делают их чрезвычайно неточными. Они также могут ввести в заблуждение относительно машинного обучения или модели больших данных. Однако выбросы иногда могут содержать полезную информацию. В результате они должны быть надлежащим образом исследованы.
7. Что вы подразумеваете под товарным оборудованием?	Товарное оборудование - это основной аппаратный ресурс, необходимый для работы платформы Apache Hadoop. Это общий термин, обозначающий доступные устройства, как правило, совместимые с другими недорогими устройствами.
8. Дайте определение и опишите FSCK.	FSCK расшифровывается как Проверка файловой системы, используемая HDFS. Она проверяет, повреждены ли какие-либо файлы, имеют ли они копии или отсутствуют какие-либо блоки. FSCK генерирует сводный отчет, который охватывает общее состояние файловой системы. Например, HDFS получает уведомление об отсутствии каких-либо файловых блоков с помощью этой команды. В отличие от обычной утилиты FSCK в Hadoop, FSCK только проверяет наличие ошибок в системе и не исправляет их.
9. Назовите номера портов для NameNode, Task Tracker и Job Tracker.	NameNode – порт 50070 Отслеживание вакансий – порт

		50030 Отслеживание задач – порт 50060
	10. Что вы понимаете под индексацией в HDFS?	HDFS индексирует блоки данных в соответствии с их размером. Конец блока данных указывает на местоположение следующего блока данных. DataNodes хранят блоки данных, в то время как NameNodes хранят эти блоки данных.
	11. Объясните переобучение в big data. Как избежать того же.	<p>переобучение - это ошибка моделирования, возникающая, когда функция жестко подгоняется под ограниченное количество точек данных. В результате получается чрезмерно сложная модель, что еще больше усложняет объяснение причуд или особенностей данных. переобучение снижает предсказуемость таких моделей. Этот эффект снижает способность к обобщению, приводя к тому, что они не могут быть обобщены при применении вне выборочных данных. Существует несколько методов избежать переобучения. Некоторые из них перечислены ниже:</p> <p>Перекрестная проверка: этот метод разбивает данные на множество небольших наборов тестовых данных, которые могут быть использованы для модификации модели.</p> <p>Регуляризация: Этот метод наказывает за все параметры, кроме перехвата, так что модель обобщает данные, а не переобучает их.</p> <p>Ранняя остановка: после определенного количества итераций способность модели к обобщению снижается; чтобы избежать этого, используется процедура, известная как ранняя остановка, для предотвращения переоснащения до того, как модель достигнет этой точки.</p>
	12. Что такое Zookeeper? Каковы преимущества использования zookeeper?	Zookeeper - это централизованное хранилище

		<p>данных, которое позволяет распределенным приложениям хранить и извлекать данные. Оно поддерживает работу разрозненной системы как единого целого, используя ее цели синхронизации, сериализации и координации.</p> <p>Способность Hadoop разделять и властвовать с зрителями зоопарков - это ее уникальный метод решения проблем с большими данными. Решение зависит от использования методов распределенной и параллельной обработки по всему кластеру Hadoop после разделения проблемы. Hadoop использует zookeeper для управления всеми компонентами этих распределенных приложений.</p> <p>У использования zookeeper есть несколько преимуществ:</p> <p>Атомарность: частичная передача данных не происходит; она либо завершается успешно, либо завершается сбоем.</p> <p>Надежность: вся система не разрушается при сбое одного узла или нескольких систем.</p> <p>Синхронизация: совместная работа и взаимное исключение между серверными процессами. Apache HBase извлекает выгоду из этого процесса для управления конфигурацией.</p> <p>Простой процесс распределенной координации: Процесс координации между всеми узлами Zookeeper прост.</p> <p>Сериализация: Сериализация - это процесс кодирования данных в соответствии с определенными правилами. Убедитесь, что ваша</p>
--	--	---

		<p>программа работает стабильно.</p> <p>Организованные сообщения: Zookeeper отслеживает сообщения с номером, отмечая их порядок отметкой о каждом обновлении; сообщения упорядочиваются здесь с помощью всего этого.</p>
	13. Объясните типы узлов Zookeeper.	<p>Узлы Zookeeper классифицируются как постоянные, эфемерные или последовательные.</p> <p>Постоянный: znode по умолчанию в zookeeper постоянно остается на сервере zookeeper, если только какие-либо другие клиенты не удалят его.</p> <p>Эфемерные: Это временные узлы zookeeper. Они удаляются, когда клиент выходит из системы с сервера ZooKeeper.</p> <p>Последовательный: Последовательные znode могут быть как эфемерными, так и постоянными. Когда новый znode создается как последовательный znode, ZooKeeper присваивает путь к znode, вставляя 10-значный порядковый номер в исходное имя.</p>
	14. Что такое MapReduce в Hadoop?	<p>MapReduce - это платформа Hadoop, используемая для обработки больших наборов данных. Другое название - модель программирования, которая позволяет нам обрабатывать большие наборы данных в компьютерных кластерах. Эта программа обеспечивает распределенное хранение данных, упрощая сложную обработку огромных объемов данных.</p> <p>Программа MapReduce работает в две разные фазы: сопоставление и сокращение.</p>

		<p>Задачи Map связаны с сопоставлением и разделением данных, в то время как задачи Reduce перетасовывают и сокращают данные.</p> <p>Hadoop может запускать приложения MapReduce на различных языках, включая Java, Ruby, Python и C ++. Программы Map Reduce в облачных вычислениях работают параллельно, что делает их идеальными для выполнения крупномасштабной обработки данных на нескольких компьютерах в кластере.</p>
	15. Когда использовать MapReduce с большими данными.	<p>MapReduce подходит для итеративных вычислений с использованием огромных объемов данных, которые должны обрабатываться параллельно. Он также подходит для крупномасштабного анализа графиков.</p>
	16. Объясните выбор функции.	<p>Большие данные могут содержать большой объем данных, которые не являются необходимыми при обработке. Таким образом, от нас может потребоваться выбрать только определенные аспекты, которые нас интересуют. Выбор функций означает извлечение только основных функций из больших данных.</p> <p>Методы выбора функций включают -</p> <p>Метод фильтров: В этом методе ранжирования переменных мы учитываем только важность и полезность функции.</p> <p>Метод оберток: В этом методе используется "алгоритм индукции", который может быть использован для создания классификатора.</p>

		Встроенный метод: Этот метод сочетает в себе преимущества как методов фильтрации, так и методов оболочки.
	17. Упомяните основные методы Reducer.	<p>Существует три основных метода редукции:</p> <p>настройка () - настройка различных параметров, таких как распределенный кэш, размер кучи и входные данные.</p> <p>reduce() - параметр, вызываемый один раз для каждого ключа с соответствующей задачей сокращения.</p> <p>очистка () - удаляет все временные файлы и выполняется только в конце задачи reducer.</p>
	18. Что такое разделение в Hive?	Разбиение на разделы в Hive означает разделение таблицы на разделы на основе значений определенного столбца, такого как дата, город, курс или страна. Затем эти разделы дополнительно подразделяются на сегменты для структурирования данных, которые могут использоваться для более эффективного выполнения запросов. Разделение может сократить время ответа на запрос, поскольку данные хранятся в виде фрагментов.
	19. Каковы параметры конфигурации в фреймворке “MapReduce”?	<p>Параметрами конфигурации в фреймворке MapReduce являются</p> <p>Укажите местоположение заданий в распределенной файловой системе</p> <p>Расположение выходных данных заданий в распределенной файловой системе</p> <p>Формат ввода данных</p> <p>Формат вывода данных</p> <p>Класс, включая функцию отображения</p> <p>Класс, включая функцию сокращения</p> <p>JAR-файл, содержащий классы Mapper, Reducer и driver.</p>

	20. Как проверяется качество данных?	<p>При тестировании больших данных качество данных так же важно, как и вычислительная мощность. База данных должна быть проверена в рамках процесса тестирования, чтобы подтвердить качество данных. Это включает в себя оценку нескольких характеристик, включая соответствие, совершенство, повторяемость, надежность, валидность, полноту данных и т.д.</p>
	21. Какие существуют типы тестирования на больших данных?	<p>типы тестирования на больших данных:</p> <p>Типы тестирования большими данными</p> <p>Функциональное тестирование: с операционными и аналитическими компонентами требует обширного функционального тестирования на уровне API. Это касается всех подкомпонентов, скриптов, программ и инструментов для хранения, загрузки и обработки прикладных приложений.</p> <p>Тестирование базы данных: Как следует из названия, это тестирование часто включает проверку данных, полученных из многочисленных баз данных. Оно гарантирует, что данные, собранные из облачных источников или локальных баз данных, являются полными и точными.</p> <p>Тестирование производительности: Автоматизация в big data помогает оценить производительность при многих обстоятельствах, таких как тестирование приложения с различными типами данных и объемами. Одним из важных методов тестирования больших данных является тестирование</p>



	<p>производительности, которое гарантирует, что задействованные компоненты обеспечивают адекватные возможности хранения, обработки и поиска больших наборов данных.</p> <p>Тестирование архитектуры: Это тестирование проверяет правильность обработки данных и соответствие бизнес-требованиям. Кроме того, если архитектура неадекватна, это может привести к снижению производительности, что приведет к прерыванию обработки данных и потере данных.</p>
22. Перечислите несколько преимуществ тестирования на больших данных.	<p>преимущества тестирования на больших данных:</p> <p>Усовершенствованный таргетинг на рынок и стратегии</p> <p>Стоимость качества</p> <p>Минимизирует потери и увеличивает доход</p> <p>Улучшенные бизнес-решения</p> <p>Точность и валидация данных</p>
23. Каковы общие проблемы при тестировании производительности?	<p>Большие данные - это комбинация нескольких технологий. Каждый подэлемент относится к отдельному оборудованию и должен тестироваться отдельно. Ниже приведены некоторые существенные проблемы, возникающие при проверке больших данных:</p> <p>Разнообразный набор технологий: каждый подкомпонент относится к другой технологии и должен тестироваться отдельно.</p> <p>Написание сценариев: Разработка тестовых примеров</p>

	<p>требует высокого уровня знаний в области написания сценариев.</p> <p>Ограниченная доступность определенных инструментов: одно устройство не может выполнить сквозное тестирование. NoSQL, например, может не подходить для очередей сообщений.</p> <p>Тестовая среда: Из-за большого объема данных требуется специализированная тестовая среда.</p> <p>Решение для мониторинга: Существует очень ограниченное количество решений для мониторинга всей среды.</p> <p>Диагностическое решение: Для разработки и устранения узкого места для повышения производительности требуется индивидуальное решение.</p>															
24. В чем разница между тестированием на больших данных и Традиционным тестированием базы данных?	<table><tr><th>Ключевые параметры</th><th>Тестирование на больших данных</th><th>Традиционное тестирование базы данных</th></tr><tr><td>Тип данных</td><td>Работает как со структурированными, так и с неструктурированными данными.</td><td>Работает только со структурированными данными.</td></tr><tr><td>Инфраструктура</td><td>Большие размеры данных и файлов (HDFS) требуют специальной тестовой среды.</td><td>не требует специальной тестовой среды, поскольку размер файла ограничен.</td></tr><tr><td>Объем данных</td><td>Его объем варьируется от петабайт до зеттабайт или эксабайт.</td><td>Его объем варьируется от гигабайт до терабайт.</td></tr><tr><td>Инструменты проверки и достоверности</td><td>Нет определенных инструментов в. Диапазон широк, от программных инструментов, таких как</td><td>Использует либо макросы на основе Excel, либо инструменты автоматизации на основе пользовательского интерфейса.</td></tr></table>	Ключевые параметры	Тестирование на больших данных	Традиционное тестирование базы данных	Тип данных	Работает как со структурированными, так и с неструктурированными данными.	Работает только со структурированными данными.	Инфраструктура	Большие размеры данных и файлов (HDFS) требуют специальной тестовой среды.	не требует специальной тестовой среды, поскольку размер файла ограничен.	Объем данных	Его объем варьируется от петабайт до зеттабайт или эксабайт.	Его объем варьируется от гигабайт до терабайт.	Инструменты проверки и достоверности	Нет определенных инструментов в. Диапазон широк, от программных инструментов, таких как	Использует либо макросы на основе Excel, либо инструменты автоматизации на основе пользовательского интерфейса.
Ключевые параметры	Тестирование на больших данных	Традиционное тестирование базы данных														
Тип данных	Работает как со структурированными, так и с неструктурированными данными.	Работает только со структурированными данными.														
Инфраструктура	Большие размеры данных и файлов (HDFS) требуют специальной тестовой среды.	не требует специальной тестовой среды, поскольку размер файла ограничен.														
Объем данных	Его объем варьируется от петабайт до зеттабайт или эксабайт.	Его объем варьируется от гигабайт до терабайт.														
Инструменты проверки и достоверности	Нет определенных инструментов в. Диапазон широк, от программных инструментов, таких как	Использует либо макросы на основе Excel, либо инструменты автоматизации на основе пользовательского интерфейса.														

		MapReduce, до HIVEQL.	
	<b>Размер данных</b>	Размер данных больше, чем у традиционны х баз данных.	Объем данных очень мал.
25. Что такое всплеск запросов?	<p>Query Surge - одно из решений для тестирования больших данных. Оно поддерживает качество данных и подход к тестированию общих данных, который обнаруживает неверные данные во время тестирования и обеспечивает отличную перспективу работоспособности данных. Это гарантирует, что данные, полученные из источников, остаются неизменными для целевого объекта, анализируя и обнаруживая различия в больших данных, когда это необходимо.</p>		
26. Какие преимущества предоставляет Query Surge?	<p>Query Surge предоставляет следующие преимущества:</p> <p>В тысячи раз увеличивает скорость тестирования, охватывая весь набор данных.</p> <p>Query Surge помогает нам автоматизировать ручное тестирование больших данных. Он тестирует несколько платформ, таких как Hadoop, Teradata, Oracle, Microsoft, IBM, MongoDB, Cloudera, Amazon и других поставщиков Hadoop.</p> <p>Он также предоставляет автоматические отчеты по электронной почте с информационными панелями, которые показывают состояние данных.</p> <p>Обеспечивает отличную отдачу от инвестиций (ROI) до 1500%.</p>		
27. Что такое тестирование на больших данных?	<p>Большие данные - это большой набор структурированных и неструктурированных данных, которые трудно обработать с</p>		

	<p>помощью традиционных баз данных и программного обеспечения. Объем данных во многих компаниях велик, и в наше время они перемещаются слишком быстро, превышая существующие вычислительные мощности. Составление баз данных, которые не могут быть эффективно обработаны традиционными компьютерными методами. Для управления этими большими объемами данных тестирование требует использования специальных инструментов, фреймворков и процессов. Анализ больших данных относится к генерации данных и их хранению, извлечению данных и анализу больших данных с точки зрения изменения объема и скорости.</p>
28. Какова цель А / В тестирования?	<p>А / В тестирование - это сравнительное исследование, в ходе которого случайным пользователям показываются две или более версий страниц, а их комментарии статистически анализируются, чтобы определить, какая версия работает лучше.</p>
29. Почему HDFS подходит только для больших наборов данных и не подходит для многих небольших файлов?	<p>Это связано с проблемой производительности NameNode. NameNode часто занимает много места для хранения метаданных для крупномасштабных файлов. Метаданные должны поступать из одного файла для оптимального использования пространства и экономической выгоды. NameNode не использует все пространство для небольших файлов, что является проблемой оптимизации производительности.</p>

## 2.2. Практико-ориентированные задания

**Практико-ориентированные задания по дисциплине «Технологии обработки больших данных» не предусмотрены.**

### 7.3. Тесты

Шифр компетенции	Тестовые задания	Правильный ответ
ПKN - 4	1. Какие из следующих технологий СУБД не используют принцип MapReduce 1) Hadoop 2) Cassandra 3) Redis 4) HDInsight	2
	2. Какие вероятные разочарования тренда больших данных? 1) из-за угрозы безопасности личной жизни (privacy) граждан будут упрощены процедуры сбора данных, что приведёт к падению ценности больших данных 2) из-за угрозы безопасности личной жизни (privacy) граждан будут усложнены процедуры сбора данных, что приведёт к падению ценности больших данных 3) нет	2
	3. Отметьте те из вариантов, в которых данные структурированы: 1) данные о продажах компании, представленные в виде ежемесячных отчётов в формате MS Word 2) библиотека фильмов, представленных в формате mpeg4 на одном жестком диске 3) таблица с ежедневными показаниями температуры помещения за год в файле формата csv 4) текст педагогической поэмы А.С. Макаренко, представленный в формате PDF	3
	4. Компания, проводящая социологические опросы получает анкеты от волонтеров, непосредственно опрашивающих респондентов. При каких условиях разумна постановка задачи цензурирования? 1) Часть анкет пришла в негодность, что не позволяет считать информацию с них со 100% уверенностью многие анкеты заполнены не полностью 2) стало известно, что волонтеры фальсифицируют результаты опроса, самостоятельно заполняя часть анкет 3) от заказчика поступило требование уничтожить часть анкет, содержащих информацию о руководителях страны	2
	5. К какому типу шкал относится шкала «очень плохо»-«плохо»-«средне»-«хорошо»-«очень хорошо»? 1) Номинальная 2) Абсолютная 3) порядковая 4) бинарная	3
	6. В чём состоит свойство расширяемости записей СУБД? 1) СУБД не имеет чёткой структуры, поэтому любую запись можно расширить 2) повышение отказоустойчивости системы при добавлении новых записей в СУБД 3) в любую таблицу СУБД можно добавить новую	4

	<p>колонку, предварительно изменив структуру этой таблицы</p> <p>4) СУБД имеет чёткую, но расширяемую структуру, в каждую запись можно добавить новую колонку, также, как и узнать значение любой записи по добавленной колонке</p>	
	<p>7. Начиная с каких размеров данных обоснованно применение кластера Hadoop для хранения данных?</p> <p>1) 100Тб</p> <p>2) 1 Пб</p> <p>3) 100Гб</p> <p>4) 1Тб</p>	<p>1</p> <p>2</p>
	<p>8. Инвестиционный фонд интересуется тем, почему часть финансируемых им проектов успешно переходят на второй год, а часть — нет. К какому типу относится эта задача анализа данных?</p> <p>1) построение решающего правила классификация</p> <p>2) поиск информативных признаков</p> <p>3) Цензурирование</p>	2
	<p>9. Компания, проводящая социологические опросы, испытывает сложности с верификацией данных, поступающих от волонтеров непосредственно опрашивающих респондентов: многие анкеты заполнены не полностью; волонтеры фальсифицируют результаты опроса, самостоятельно заполняя часть анкет. К какому типу задач анализа данных здесь прибегать не придётся?</p> <p>1) Классификация</p> <p>2) Цензурирование</p> <p>3) прогнозирование</p> <p>4) заполнение пробелов</p>	3
	<p>10. На каком из этапов процесса CRISP-DM происходит проверка гипотез?</p> <p>1) моделирование (Modeling)</p> <p>2) понимание данных (Data Understanding)</p> <p>3) оценка (Evaluation)</p> <p>4) понимание бизнеса (Business understanding)</p>	1

## 8.Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

### Основная литература:

1. Колдаев, В. Д. Структуры и алгоритмы обработки данных : учебное пособие / В. Д. Колдаев. - Москва : РИОР : ИНФРА-М, 2021. - 296 с. - ЭБС ZNANIUM.com. - URL: <https://znanium.com/catalog/product/1230215> (дата обращения: 07.12.2024). – Текст : электронный.
2. Баланов, А. Н. Цифровое понимание. Создание, влияние и будущее технологий : учебник для вузов / А. Н. Баланов. Санкт-Петербург : Лань, 2024. - 452 с. - ISBN 978-5-507-49416-3. - Текст: электронный // Лань : электронно-библиотечная система. - URL:<https://e.lanbook.com/book/417800> (дата обращения:19.07.2024).
3. Коломейченко, А. С. Информационные технологии : учебное пособие для спо / А. С. Коломейченко, Н. В.Польшакова, О. В. Чеха. - 3-е изд., стер. - Санкт-Петербург : Лань, 2024. - 212 с. - ISBN 978-5-507-49263-3. - Текст : электронный // Лань : электронно-

библиотечная система. - URL: <https://e.lanbook.com/book/384743> (дата обращения: 19.07.2024).

#### **Дополнительная литература:**

4. Нагаева, И. А. Основы алгоритмизации и программирования: практикум : учебное пособие / И. А. Нагаева, И. А. Кузнецов. – Москва : Берлин : Директ-Медиа, 2021. – 169 с. – ЭБС Университетская библиотека ONLINE. – URL: <https://biblioclub.ru/index.php?page=book&id=598404> (дата обращения: 07.12.2024). – Текст : электронный.

5. Баланов, А. Н. Big Data и анализ статистики в спорте : учебное пособие для вузов / А. Н. Баланов. - Санкт-Петербург : Лань, 2024. - 272 с. - ISBN 978-5-507-49244- Текст : электронный // Лань : электронно-библиотечная система. - URL: <https://e.lanbook.com/book/414875> (дата обращения: 19.07.2024).

### **9.Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины**

1. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/>
2. Электронно-библиотечная система BOOK.RU <http://www.book.ru>
3. Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» <http://biblioclub.ru/>
4. Электронно-библиотечная система Znanium <http://www.znanium.comPylru> 1.0.9 [Электронный ресурс]: сайт. – Режим доступа: <https://pypi.python.org/pypi/pylru>
5. Python Data Analysis Library [Электронный ресурс]: сайт. – Режим доступа: <http://pandas.pydata.org/>
6. Python Documentation [Электронный ресурс]: сайт. – Режим доступа: <http://python.org/doc/>
7. Python Standard Library [Электронный ресурс]: сайт. – Режим доступа: <https://docs.python.org/2/library/>
8. Scikit-learn Machine Learning in Python [Электронный ресурс]: сайт. – Режим доступа: <http://scikit-learn.org>
9. Официальный сайт продукта <https://www.python.org/>
10. Каталог курсов Интернет Университета Информационных Технологий <http://www.intuit.ru/>
11. The Python Tutorial // <https://docs.python.org/3/tutorial/index.html>
12. NumPy User Guide // <http://docs.scipy.org/doc/numpy/user/index.html>
13. Pandas User Guide <http://pandas.pydata.org/pandas-docs/stable/>
14. Dask User Guide <https://docs.dask.org/en/latest/>
15. Dask User Guide <https://docs.dask.org/en/latest/>
16. Matplotlib User Guide // <https://matplotlib.org/stable/users/index.html>
17. Seaborn User Guide // <https://seaborn.pydata.org/tutorial.html>

### **10.Методические указания для обучающихся по освоению дисциплины**

При изучении теоретического материала необходимо опираться на рабочую программу дисциплины, материалы лекций и литературу из основного списка. Кроме

этого, необходимо активно работать с Интернет-источниками и пособиями других авторов, помогающими усвоить материал отдельных разделов программы.

Необходимо конспектировать лекции, помечая сложные и непонятные моменты с тем, чтобы задать вопросы лектору в конце лекции или же на консультации.

При подготовке к семинарским занятиям необходимо изучить вопросы, вынесенные на самостоятельное изучение, так как семинарские занятия предполагают их обсуждение и дискуссию по теме; кроме того, задания для самостоятельной работы необходимы для того, чтобы успешно выполнить самостоятельные задания на семинарах.

Индивидуальные задания для работы на компьютере, файлы с выполненными заданиями необходимо хранить в личной сетевой папке в компьютерной сети вуза.

### **11.Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем**

11.1. Комплект лицензионного программного обеспечения:

1. Пакет офисных программ

2. Антивирус Kaspersky

11.2. Современные профессиональные базы данных и информационные справочные системы

1. Информационно-правовая система «Гарант»

2. Информационно-правовая система «Консультант Плюс»

3. Электронная энциклопедия: <http://ru.wikipedia.org/wiki/Wiki>

4. Система комплексного раскрытия информации «СКРИН» - <http://www.skrin.ru/>

11.3. Сертифицированные программные и аппаратные средства защиты информации

- не используются

### **12.Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине**

Для проведения лекций и практических занятий необходима аудитория, оснащенная проектором и компьютерами с постоянным подключением к сети Интернет.