

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

38.03.05 – «Бизнес-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: Кирий В.А., кандидат физ.-мат. наук, доцент кафедры «Математика и информатика», Арефьева С.А., кандидат техн.наук, доцент, доцент кафедры «Математика и информатика»

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 .....	24
7.3.Пример экзаменационного билета .....	29
8.Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины .....	31
9.Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины .....	32
10.Методические указания для обучающихся по освоению дисциплины .....	33
11.Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем .....	33
12.Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине .....	33

## 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><u>Знать:</u></b> методы и инструменты анализа данных и машинного обучения <b><u>Уметь:</u></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	5/144	144
Контактная работа - Аудиторные занятия	28	28
Лекции	12	12
Семинарские, практические занятия	16	16
Самостоятельная работа	116	116
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
		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		144	28	12	16	116	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.	Продвинутые операции 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»	
ПКП-3 Способность предлагать различные варианты инфраструктурных решений для поддержки ИТ/ИС					
Анализирует текущий уровень инфраструктурных решений предприятия/организации.					
Знать: методы и подходы к обследованию и анализу текущего уровня инфраструктурных решений предприятия.	Фрагментарное представление о методах и подходах к обследованию и анализу текущего уровня инфраструктурных решений предприятия.	Неполные представления о методах и подходах к обследованию и анализу текущего уровня инфраструктурных решений предприятия.	Сформированные, но содержащие отдельные пробелы представления о методах и подходах к обследованию и анализу текущего уровня инфраструктурных решений предприятия.	Сформированные систематические представления о методах и подходах к обследованию и анализу текущего уровня инфраструктурных решений предприятия.	Вопросы для оценки знаний и умений, тестовые задания
Уметь: проводить анализ и формулировать обоснованные выводы о текущем уровне инфраструктурных решений организации	Фрагментарное умение проводить анализ и формулировать обоснованные выводы о текущем уровне инфраструктурных решений организации	Несистематическое применение умений проводить анализ и формулировать обоснованные выводы о текущем уровне инфраструктурных решений организации	В целом успешное, но содержащее отдельные пробелы умение проводить анализ и формулировать обоснованные выводы о текущем уровне инфраструктурных решений организации	Сформированное умение проводить анализ и формулировать обоснованные выводы о текущем уровне инфраструктурных решений организации	Вопросы для оценки знаний и умений, тестовые задания
Формирует и обосновывает варианты технологического слоя архитектуры предприятия/организации.					
Знать:	Фрагментарное	Неполные	Сформирован	Сформирован	Вопрос

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»	
рекомендации TOGAF по анализу и моделированию технологического слоя архитектуры предприятия	представление о рекомендациях TOGAF по анализу и моделированию технологического слоя архитектуры предприятия	представления о рекомендациях TOGAF по анализу и моделированию технологического слоя архитектуры предприятия	ные, но содержащие отдельные пробелы представления о рекомендациях TOGAF по анализу и моделированию технологического слоя архитектуры предприятия	ные систематические представления о рекомендациях TOGAF по анализу и моделированию технологического слоя архитектуры предприятия	ы для оценки знаний и умений, тестовые задания
Уметь: разрабатывать артефакты текущего и целевого представления технологического слоя архитектуры предприятия	Фрагментарное умение разрабатывать артефакты текущего и целевого представления технологического слоя архитектуры предприятия	Несистематическое применение умений разрабатывать артефакты текущего и целевого представления технологического слоя архитектуры предприятия	В целом успешное, но содержащее отдельные пробелы умение разрабатывать артефакты текущего и целевого представления технологического слоя архитектуры предприятия	Сформированное умение разрабатывать артефакты текущего и целевого представления технологического слоя архитектуры предприятия	Вопросы для оценки знаний и умений, тестовые задания
<b>ПКН-9 Способность управлять моделью сорсинга</b>					
Демонстрирует знания о моделях сорсинга					
Знать: модели сорсинга	Фрагментарное представление о моделях сорсинга	Неполные представления о моделях сорсинга	Сформированные, но содержащие отдельные пробелы представления о моделях сорсинга	Сформированные систематические представления о моделях сорсинга	Вопросы для оценки знаний и умений, тестовые задания
Уметь: использовать технологии, услуги и	Фрагментарное умение использовать технологии,	Несистематическое применение умений использовать	В целом успешное, но содержащее отдельные	Сформированное умение использовать технологии,	Вопросы для оценки знаний

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»	
ресурсы как внешних, так и внутренних ИТ-провайдеров для решения бизнес-задач.	услуги и ресурсы как внешних, так и внутренних ИТ-провайдеров для решения бизнес-задач.	технологии, услуги и ресурсы как внешних, так и внутренних ИТ-провайдеров для решения бизнес-задач.	пробелы умение использовать технологии, услуги и ресурсы как внешних, так и внутренних ИТ-провайдеров для решения бизнес-задач.	услуги и ресурсы как внешних, так и внутренних ИТ-провайдеров для решения бизнес-задач.	и умений, тестовые задания
<b>Применяет различные модели сорсинга для конкретных предприятий.</b>					
Знать: какие критерии в первую очередь влияют на выбор целевой формы сорсинга ИТ	Фрагментарное представление о критериях в первую очередь влияющих на выбор целевой формы сорсинга ИТ	Неполные представления о критериях в первую очередь влияющих на выбор целевой формы сорсинга ИТ	Сформированные, но содержащие отдельные пробелы представления о критериях в первую очередь влияющих на выбор целевой формы сорсинга ИТ	Сформированные систематическое представления о критериях в первую очередь влияющих на выбор целевой формы сорсинга ИТ	Вопросы для оценки знаний и умений, тестовые задания
Уметь: Выбирать целевые формы сорсинга ИТ	Фрагментарное умение выбора целевых форм сорсинга ИТ	Несистематическое применение умений выбора целевых форм сорсинга ИТ	В целом успешное, но содержащее отдельные пробелы умение выбора целевых форм сорсинга ИТ	Сформированное умение выбора целевых форм сорсинга ИТ	Вопросы для оценки знаний и умений, тестовые задания
<b>ПКН-10 Способность применять знания по сервисноориентированному подходу в ИТ и консультировать по вопросам управления ИТ-сервисами</b>					
<b>Проектирует каталог ИТ-услуг.</b>					
Знать: принципы построения каталога ИТ - услуг.	Фрагментарное представление о принципах построения каталога ИТ - услуг	Неполные представления о принципах построения каталога ИТ - услуг	Сформированные, но содержащие отдельные пробелы представления о принципах	Сформированные систематическое представления о принципах построения	Вопросы для оценки знаний и умений, тестовые

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»	
			построения каталога ИТ - услуг	каталога ИТ - услуг	е задания
Уметь: систематизировать ИТ - услуги компании и формировать соответствующий каталог	Фрагментарное умение систематизировать ИТ - услуги компании и формировать соответствующий каталог	Несистематическое применение умений систематизировать ИТ - услуги компании и формировать соответствующий каталог	В целом успешное, но содержащее отдельные пробелы умение систематизировать ИТ - услуги компании и формировать соответствующий каталог	Сформированное умение систематизировать ИТ - услуги компании и формировать соответствующий каталог	Вопросы для оценки знаний и умений, тестовые задания
Выявляет ИТ-процессы, необходимые для реализации ИТ-сервисов.					
Знать: современные классификации ИТ-процессов и ИТ - сервисов, представленные в ITIL и COBIT.	Фрагментарное представление о современных классификациях ИТ-процессов и ИТ - сервисов, представленных в ITIL и COBIT.	Неполные представления о современных классификациях ИТ-процессов и ИТ - сервисов, представленных в ITIL и COBIT.	Сформированные, но содержащие отдельные пробелы представления о современных классификациях ИТ-процессов и ИТ - сервисов, представленных в ITIL и COBIT.	Сформированные систематические представления о современных классификациях ИТ-процессов и ИТ - сервисов, представленных в ITIL и COBIT.	Вопросы для оценки знаний и умений, тестовые задания
Уметь: применять COBIT и ITIL для определения ИТ - процессов, необходимых для реализации ИТ- сервисов.	Фрагментарное умение применять COBIT и ITIL для определения ИТ - процессов, необходимых для реализации ИТ- сервисов.	Несистематическое умение применять COBIT и ITIL для определения ИТ - процессов, необходимых для реализации ИТ- сервисов.	В целом успешное, но содержащее отдельные пробелы умение применять COBIT и ITIL для определения ИТ - процессов, необходимых	Сформированное умение применять COBIT и ITIL для определения ИТ - процессов, необходимых для реализации ИТ- сервисов.	Вопросы для оценки знаний и умений, тестовые задания

Планируемые результаты освоения компетенции (индикатора достижения компетенции)	Уровень освоения				Оценочное средство
	«неудовлетворительно»	«удовлетворительно»	«хорошо»	«отлично»	
			для реализации ИТ- сервисов.		

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

Шифр компетенции	Вопросы	Правильный ответ			
ПКП-3	1. Дайте определение большим данным и объясните семь Против больших данных.	<p>Большие данные - это набор больших и сложных полуструктурированных и неструктурированных наборов данных, которые потенциально могут предоставлять полезную информацию с использованием традиционных инструментов управления данными.</p> <p>Семь Против больших данных - это</p> <p>Объем: Объем представляет собой объем данных, растущий экспоненциально.</p> <p>Скорость: Скорость представляет собой скорость, с которой растут данные.</p> <p>Разнообразие: Разнообразие относится к типам данных в различных форматах данных.</p> <p>Ценность: Ценность означает получение ценной информации для удовлетворения потребностей бизнеса и получения доходов.</p> <p>Правдивость: Правдивость связана с точностью 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>			
	2. Как развернуть модель больших данных? Укажите ключевые шаги.	<p>Развертывание модели больших данных включает в себя три этапа:</p> <ol style="list-style-type: none"> <li>1. Прием данных</li> <li>2. Хранение данных</li> <li>3. Обработка данных</li> </ol>			
	3. Как Hadoop связан с большими данными?	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	<b>Ключевые особенности</b>	<b>Hadoop</b>	<b>СУБД</b>	



	Hadoop и RDBMS.	Обзор	Hadoop - это набор программного обеспечения с открытым исходным кодом, который объединяет несколько компьютеров для решения задач, требующих больших объемов данных и их обработки.	СУБД - это часть программного обеспечения, используемого для хранения данных на основе реляционной модели и управления ими.	
		Разнообразие данных	Hadoop хранит структурированные, полуструктурированные и неструктурированные данные.	СУБД хранит структурированные данные.	
		Хранение данных	Hadoop хранит большие наборы данных.	В СУБД хранятся структурированные данные.	
		Аппаратное обеспечение	Hadoop использует обычное оборудование.	СУБД использует высокопроизводительные серверы.	
		Масштабируемость	Hadoop обладает горизонтальной масштабируемостью.	СУБД обладает вертикальной масштабируемостью.	
		Пропускная способность	Высокий	Низкий уровень	
	5. Назовите несколько методов обработки больших данных.	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.			
	7. Что вы подразумеваете под товарным оборудованием ?	Товарное оборудование - это основной аппаратный ресурс, необходимый для работы платформы Apache Hadoop.			
	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.			
	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.			

ПКН -9	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.</p> <p>Существует несколько методов избежать переобучения:</p> <ol style="list-style-type: none"> <li>1. Перекрестная проверка</li> <li>2. Регуляризация</li> <li>3. Ранняя остановка</li> </ol>
	12. Что такое Zookeeper? Каковы преимущества использования zookeeper?	<p>Zookeeper - это централизованное хранилище 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> <ol style="list-style-type: none"> <li>1. Атомарность</li> <li>2. Надежность</li> <li>3. Синхронизация</li> <li>4. Простой процесс распределенной координации</li> <li>5. Сериализация</li> <li>6. Организованные сообщения</li> </ol>
	13. Объясните типы узлов Zookeeper.	<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.</p> <p>Программа MapReduce работает в две разные фазы: сопоставление и сокращение. Задачи Map связаны с сопоставлением и разделением данных, в то время как задачи Reduce перетасовывают и сокращают данные.</p>
	15. Когда	MapReduce подходит для итеративных вычислений с

	использовать MapReduce с большими данными.	использованием огромных объемов данных, которые должны обрабатываться параллельно.
	16. Объясните выбор функции.	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.
	17. Упомяните основные методы Reducer.	настройка () - настройка различных параметров, таких как распределенный кэш, размер кучи и входные данные. reduce() - параметр, вызываемый один раз для каждого ключа с соответствующей задачей сокращения. очистка () - удаляет все временные файлы и выполняется только в конце задачи reducer.
	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.
	19. Каковы параметры конфигурации в фреймворке "MapReduce"?	Укажите местоположение заданий в распределенной файловой системе Расположение выходных данных заданий в распределенной файловой системе Формат ввода данных Формат вывода данных Класс, включая функцию отображения Класс, включая функцию сокращения JAR-файл, содержащий классы Mapper, Reducer и driver.
	20. Как проверяется качество данных?	Включает в себя оценку нескольких характеристик, включая соответствие, совершенство, повторяемость, надежность, валидность, полноту данных и т.д.
ПКН -10	21. Какие существуют типы тестирования на больших данных?	Функциональное тестирование: с операционными и аналитическими компонентами требует обширного функционального тестирования на уровне API. Тестирование базы данных: Как следует из названия, это тестирование часто включает проверку данных, полученных из многочисленных баз данных. Тестирование производительности: Автоматизация в big data помогает оценить производительность при многих обстоятельствах, таких как тестирование приложения с различными типами данных и объемами. Тестирование архитектуры: Это тестирование проверяет правильность обработки данных и соответствие бизнес-требованиям.
	22. Перечислите несколько преимуществ тестирования на больших данных.	Усовершенствованный таргетинг на рынок и стратегии Стоимость качества Минимизирует потери и увеличивает доход Улучшенные бизнес-решения Точность и валидация данных
	23. Каковы	Разнообразный набор технологий

	<p>общие проблемы при тестировании производительности?</p>	<p>Написание сценариев Ограниченная доступность определенных инструментов Тестовая среда Решение для мониторинга Диагностическое решение</p>			
	<p>24. В чем разница между тестированием на больших данных и Традиционным тестированием базы данных?</p>	<p><b>Ключевые параметры</b></p>	<p><b>Тестирование на больших данных</b></p>	<p><b>Традиционное тестирование базы данных</b></p>	
		<p><b>Тип данных</b></p>	<p>Работает как со структурированными, так и с неструктурированными данными.</p>	<p>Работает только со структурированными данными.</p>	
		<p><b>Инфраструктура</b></p>	<p>Большие размеры данных и файлов (HDFS) требуют специальной тестовой среды.</p>	<p>не требует специальной тестовой среды, поскольку размер файла ограничен.</p>	
		<p><b>Объем данных</b></p>	<p>Его объем варьируется от петабайт до зеттабайт или эксабайт.</p>	<p>Его объем варьируется от гигабайт до терабайт.</p>	
		<p><b>Инструменты проверки достоверности</b></p>	<p>Нет определенных инструментов. Диапазон широк, от программных инструментов, таких как MapReduce, до HIVEQL.</p>	<p>Использует либо макросы на основе Excel, либо инструменты автоматизации на основе пользовательского интерфейса.</p>	
		<p><b>Размер данных</b></p>	<p>Размер данных больше, чем у традиционных баз данных.</p>	<p>Объем данных очень мал.</p>	
	<p>25. Что такое всплеск запросов?</p>	<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.</p>			
	<p>26. Какие преимущества предоставляет Query Surge?</p>	<p>Query Surge предоставляет следующие преимущества: В тысячи раз увеличивает скорость тестирования, охватывая весь набор данных. 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. Он также предоставляет автоматические отчеты по электронной почте с информационными панелями, которые показывают состояние данных. Обеспечивает отличную отдачу от инвестиций (ROI) до 1500%.</p>			
	<p>27. Что такое тестирование</p>	<p>Большие данные - это большой набор структурированных и неструктурированных данных, которые трудно обработать с</p>			

	на больших данных?	помощью традиционных баз данных и программного обеспечения.
	28. Какова цель А / В тестирования?	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.
	29. Почему HDFS подходит только для больших наборов данных и не подходит для многих небольших файлов?	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.

## 2.2. Практико-ориентированные задания

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

### 7.3. Тесты

Шифр компетенции	Вопросы	Правильный ответ
ПКП-3	1. Data Mining – это процесс обнаружения в сырых данных знаний, необходимых для: а) расчета целевых показателей; б) принятия решений в различных сферах человеческой деятельности; с) определения области допустимых значений.	b
	2. Атрибут – это: а) мера оценки; б) значение переменной; с) свойство, характеризующее объект	c
	3. В процессе работы Data Mining программы пользователь может получить такие результаты: а) только статистически достоверные результаты; б) полученные на основе методов математического анализа; с) расчетные показатели процесса.	a
	4. Данные могут быть получены в результате: а) наблюдений; б) экспериментов; с) расчетов.	a,b
	5. Данные представляют собой: а) факты; б) наблюдения; с) меры оценки.	a,b
	6. Данные – это ...	a

	a) необработанный материал, предоставляемый поставщиками данных и используемый потребителями для формирования информации на основе данных; b) расчетные характеристики процесса; c) результаты обработки имеющихся параметров.	
	7. Два основных типа переменных в статистике: a) качественные и количественные; b) расчетные и априорные; c) простые и сложные.	a
ПКН-9	8. Определите для какой шкалы применимы только такие операции как равно и не равно. a) порядковая шкала; b) регрессивная шкала; c) номинальная шкала; d) прогрессивная шкала.	c
	9. Определите для какой шкалы применимы только такие операции как равно, не равно, больше, меньше. a) порядковая шкала; b) регрессивная шкала; c) номинальная шкала; d) прогрессивная шкала	a
	10. Задачей классификации можно назвать предсказание... a) категориальной зависимой переменной, основываясь на выборке непрерывных и/или категориальных переменных; b) категориальной независимой переменной, основываясь на выборке непрерывных и/или категориальных переменных.	a
	11. Задачей регрессии можно назвать предсказание... a) числовой зависимой переменной, основываясь на выборке непрерывных и/или категориальных переменных; b) текстовой зависимой переменной, основываясь на выборке непрерывных и/или категориальных переменных; c) числовой независимой переменной, основываясь на выборке непрерывных и/или категориальных переменных	a
	12. Задачи классификации решаются следующими алгоритмами: a) нейронные сети b) линейной регрессии c) Дейкстры d) Отжига	a,b
	13. Задачу классификации нельзя решить с помощью... a) алгоритма Apriori; b) нейронной сетью; c) алгоритма линейной регрессии	a
	14. Закономерности, найденные в процессе использования технологии Data Mining не должны обладать такими свойствами: a) быть практически полезными b) быть объективными c) быть неочевидными	d

	d) быть точными	
ПКН-10	15. Изначальная предопределенность классов является характеристикой задачи ... a) кластеризации; b) классификации; c) интеграции	b
	16. Инструменты Data Mining a) работают на основе точных методов; b) могут самостоятельно строить гипотезы о взаимосвязях в данных; c) позволяют найти глобальный оптимум	b
	17. Множество примеров, используемое для конструирования модели, называется... a) обучающим множеством; b) тестовым множеством; c) эталонным множеством; d) предметным множеством	a
	18. Множество примеров, используемое для проверки работы сконструированной модели, называется... a) обучающим множеством; b) тестовым множеством; c) эталонным множеством; d) предметным множеством	b
	19. Назовите факторы, обусловившие возникновение и развитие Data Mining: a) совершенствование алгоритмов обработки информации b) накопление большого количества ретроспективных данных c) совершенствование технологий хранения и записи данных d) совершенствование аппаратного и программного обеспечения e) появление методов математического анализа	b
	20. Номинальная шкала – это шкала, a) содержащая только категории, которые не могут упорядочиваться; b) содержащая только категории, которые могут упорядочиваться; c) содержащая только категории, которые могут интегрироваться	a
	21. Объект описывается как ... a) набор атрибутов; b) набор переменных; c) набор правил; d) набор характеристик	a

**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 978-5-507-49244-4. - 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.com/Pylru> 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 до- ступа: <http://python.org/doc/>
7. Python Standard Library [Электронный ресурс]: сайт. – Режим до- ступа: <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 Техно- логий <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.