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

Департамент анализа данных и машинного обучения
Факультета информационных технологий и анализа больших данных

СОГЛАСОВАНО	УТВЕРЖДАЮ
Газпромбанк	Проректор по учебной
Начальник отдела управления	и методической работе
алгоритмов машинного обучения	_____Е.А. Каменева
15.12.2022 г.	29.12.2022 г.

Макрушин С.В., Блохин Н.В.

ТЕХНОЛОГИИ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки
09.03.03 - Прикладная информатика,
ОП «Инженерия данных»,
Профиль: «Инженерия данных»

Рекомендовано Ученым советом
Факультета информационных технологий и анализа больших данных
(протокол №27 от 15.12.2022г.)

Одобрено Советом учебно-научного
Департамента анализа данных и машинного обучения
(протокол №6 от 13.12.2022 г.)

Москва 2022

Оглавление

1. Наименование дисциплины

«Технологии обработки больших данных».

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

Код компетенции	Наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
ПКП-2	Способность разрабатывать, согласовывать и управлять исполнением технического задания и технического проекта с использованием технологий больших данных	Работает со стандартами, в том числе адаптирует стандарты для специфических требований больших данных.	**Знать** существующие стандарты, необходимые для создания технического задания и технического проекта с учетом специфических требований больших данных **Уметь** использовать и адаптировать существующие стандарты с учетом специфических требований больших данных
		Разрабатывает технические задания и технические проекты для технологий больших данных.	**Знать** технологию разработки технических заданий и технических проектов, в которых используются технологии больших данных **Уметь** разрабатывать технические задания и технических проекты, в которых используются технологии больших данных
		Реализует управление рабочими проектами технологической инфраструктуры больших данных.	**Знать** современные принципы управления рабочими проектами, применяемыми к технологической инфраструктуре больших данных **Уметь** применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных

3. Место дисциплины в структуре образовательной программы

Дисциплина «Технологии обработки больших данных» является дисциплиной профиля «Инженерия данных» по направлению подготовки 09.03.03 – Прикладная информатика, ОП «Инженерия данных».

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

Вид учебной работы по дисциплине	Всего (в з.е. и часах)	Семестр 5 (в часах)
Общая трудоёмкость дисциплины (в том числе курсовой проект)	5/180 (курсовой проект, 24 часа)	180 (курсовой проект, 24 часа)
Контактная работа-Аудиторные занятия	***50***	***50***
Лекции	16	16
Семинары, практические занятия	34	34
Самостоятельная работа	***130***	***130***
Вид текущего контроля	Экзамен	Экзамен
Вид промежуточной аттестации	Контрольная работа	Контрольная работа

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 едиными 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, стеммминг и лемматизация, примеры на Python. Использование мемоизации на примере работы со строками. Расстояние Левеншнтейна: 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 дан-

ных 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

№ п/п	Наименование темы (раздела) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемости
		Всего	Контактная работа-Аудиторная работа			Самостоятельная работа	
			Общая, в т.ч.:	Лекции	Семинары, практические занятия		
1	Библиотека NumPy и Pandas	22	6	2	4	16	Участие в решении задач на практических занятиях. Обсуждения по результатам самостоятельной работы
2	Использование различных форматов файлов в задачах обработки данных.	22	6	2	4	16	
3	Взаимодействие с табличными данными в приложениях обработки данных.	11	3	1	2	8	
4	Визуализация данных	11	3	1	2	8	
5	Работа со строками в приложениях обработки данных	11	3	1	2	8	Участие в решении задач на практических занятиях. Обсуждения по результатам самостоятельной работы
6	Введение в обработку текста на естественном языке в задачах обработки данных	11	3	1	2	8	
7	Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba.	11	3	1	2	8	
8	Взаимодействие с базой данных в	11	3	1	2	8	

	приложениях обработки данных.						Участие в решении задач на практических занятиях. Обсуждения по результатам самостоятельной работы
9	Параллельная обработка данных	15	5	1	4	10	
10	Библиотека Dask	44	12	4	8	32	
11	Обзор проблем обработки больших данных и вычисления общего назначения на GPU	11	3	1	2	8	
	В целом по дисциплине	180	50	16	34	130	Согласно учебному плану: контрольная работа
	Итого в %		28	32	68	72	

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], 9[9], 9[10]	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.	Интерактивная форма, работа на компьютере

	• Формат XML и модель DOM: общая характеристика, пример описания данных в XML и DOM • Работа с XML с помощью библиотеки BeautifulSoup. 8[1], 8[2], 9[3], 9[4]	
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], 8[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[13], 9[15], 9[16]	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], 8[2], 9[4]	Интерактивная форма, работа на компьютере
Введение в обработку текста на естественном языке в задачах обработки данных.	• Сегментация и токенезация текста на естественном языке, стеммминг и лемматизация, примеры на Python. • Расстояние Левеншнтейна: определение, алгоритм эффективного поиска оптимального редакционного предписания, пример поиска на Python. 8[1], 8[2], 9[4], 9[5], 9[6]	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. 8[1], 8[2], 9[1], 9[2], 9[3]	Интерактивная форма, работа на компьютере
Взаимодействие с базой данных приложениях обработки данных	• Взаимодействие из Python с базой данных с помощью API SQLite. 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.	Интерактивная форма, работа на компьютере

	8[1], 8[2], 9[5], 9[7]	
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. • 8[1], 8[2], 9[8], 9[10], 9[11]	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. • 8[1], 8[2], 9[8]	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	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналити-

	ним в Python.	ческой 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	Обзор литературы и веб-источников. Самостоятельное освоение

	данных с помощью Dask.Bag. • Организация вычислений с помощью API Dask Delayed.	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 DaskBag
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 содержатся в соответствующих методических рекомендациях Департамента анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных.

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

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

Наименование компетенции	Наименование индикаторов достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции	Типовые контрольные задания

ПКП-2 Способность разрабатывать, согласовывать и управлять исполнением технического задания и технического проекта с использованием технологий больших данных	Работает со стандартами, в том числе адаптирует стандарты для специфических требований больших данных.	**Знать** существующие стандарты, необходимые для создания технического задания и технического проекта с учетом специфических требований больших данных **Уметь** использовать и адаптировать существующие стандарты с учетом специфических требований больших данных	Перечислить основные стандарты создания технического задания и технического проекта с учетом специфики больших данных Адаптируйте любой стандарт для реализации проекта с предполагаемым объемом данных не менее 1 ТБ.
	Разрабатывает технические задания и технические проекты для технологий больших данных.	**Знать** технологию разработки технических заданий и технических проектов, в которых используются технологии больших данных **Уметь** разрабатывать технические задания и технических проекты, в которых используются технологии больших данных	Описать технологию создания технического задания и технического проекта с учетом специфики больших данных Разработайте техническое задание для реализации проекта с предполагаемым объемом данных не менее 1 ТБ.
	Реализует управление рабочими проектами технологической инфраструктуры больших данных.	**Знать современные** принципы управления рабочими проектами, применяемыми к технологической инфраструктуре больших данных **Уметь** применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных	Перечислить основные принципы управления рабочими проектами, в которых задействована технологическая инфраструктура больших данных Продемонстрируйте современные принципы управления рабочими проектами на примере создания рабочей группы для разворачивания технологической инфраструктуры больших данных

Примерные вопросы для подготовки к экзамену

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. Профилирование реализации алгоритмов на Python, принципы решения задачи оптимизации производительности алгоритма
8. Проблема Global Interpreter Lock в Python и способы обхода ее ограничений
9. Технологический стек Python для обработки и анализа данных, Python как glue language, специфика библиотеки NumPy и ее роль в экосистеме Python
10. Организация массивов в NumPy: хранение данных, создание массивов, принципы реализации операций с едиными исходными данными
11. Универсальные функции и применение функций по осям в NumPy
12. Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры
13. Маскирование и прихотливое индексирование в NumPy

14. Векторизация в numpy: ключевые параметры функции, примеры применения, использование обобщенной сигнатуры функции
15. Numba: принципы работы, базовые примеры использования
16. Организация Pandas DataFrame и организация индексации для DataFrame и Series
17. Применение универсальных функций и работа с пустыми значениями в Pandas
18. Объединение данных из нескольких Pandas DataFrame: общая логика и примеры
19. Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение»
20. Специфика текстовых и бинарных файлов, форматы файлов CSV и Pickle, представление данных в этих форматах и взаимодействие с ними в Python
21. Задача сериализации и десериализации, описание формата файла JSON и пример описания данных в этом формате и взаимодействия с ним в Python
22. Формат XML и модель DOM: общая характеристика, пример описания данных в XML и DOM, работа с ними с помощью библиотеки BeautifulSoup
23. Форматы файлов NPY и HDF общая характеристика, пример взаимодействие с данными этих форматов в Python
24. Взаимодействие из Python с базой данных на примере API SQLite, базовые возможности работы с транзакциями
25. Взаимодействие с Excel из Python с помощью XLWings: принципы работы и примеры использования
26. Основы работы с регулярными выражениями: базовый синтаксис, примеры использования модуля re в Python

27. Сегментация и токенезация текста на естественном языке, стемминг и лемматизация, примеры на Python
28. Расстояние Левеншнтейна: определение, алгоритм эффективного поиска оптимального редакционного предписания, пример поиска на Python
29. Векторное представление текста на естественном языке: общий алгоритм подходов TF; TF-IDF
30. Модуль multiprocessing – назначение и основные возможности, API multiprocessing.Pool
31. Различия между потоками и процессами, различие между различными планировщиками в Dask
32. Граф зависимостей задач – суть структуры данных, ее построение и использование в Dask
33. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач
34. Dask.Array – структура данных, специфика реализации и применения, процедура создания
35. Dask.Array – поддерживаемые операции и отличия от NumPy ndarray
36. Распараллеливание алгоритмов с помощью dask.delayed – принцип и примеры использования
37. Дополнительные параметры декоратора dask.delayed – назначение и примеры использования
38. Использование dask.delayed для объектов и операции над объектами dask.delayed, включая ограничения их использования
39. Dask.DataFrame - структура данных, специфика реализации и применения, процедура создания Dask.DataFrame
40. Ограничения использования Dask.DataFrame и операции мэппинга в Dask.DataFrame

41. Поддержка Dask.DataFrame операций работающих со скользящим окном
42. Совместное использование промежуточных результатов в Dask: принцип работы и примеры использования
43. Dask.Bag - структура данных, специфика реализации и применения, процедура создания DaskBag
44. Организация вычислений с помощью Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных в Dask.Bag
45. API Dask.Bag – функции мэппинга, фильтрации и преобразования
46. API Dask.Bag – функции группировки и свертки
47. Организация системы координат в библиотеке Matplotlib
48. Понятие признака в анализе данных и типы признаков
49. Понятие разведочного анализа данных, основные задачи и типовые визуализации для решения этих задач

Пример экзаменационного билета

1. (20 баллов) Большие данные – определение и причины возникновения задач обработки больших данных.
2. (20 баллов) Датасет: Chinook_Sqlite.sqlite

   С помощью кода на Python с использованием sqlite3 и SQL решить задачу. Реализовать функции на Python:

   1. Которая возвращает все имеющиеся плейлисты.

   2. Которая по имени плейлиста возвращает количество треков в нем и их суммарную продолжительность.
3. (20 баллов) Датасет: all_k.zip

   Подсчитать, сколько раз во всех текстовых файлах, лежащих в all_k.zip, встречаются реплики прямой речи, оформленные в виде диалога (В этом случае каждая реплика начинается с новой строки, перед

репликами ставится тире (перед тире возможны различные пробельные символы)). Выполнить задание с использованием Dask, распараллелив процесс обработки данных.

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

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: 07.12.2022). – Текст : электронный.

Дополнительная литература:

2. 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: 07.12.2022). – Текст : электронный.

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 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.