

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

**Кафедра анализа данных и машинного обучения
Факультета информационных технологий и анализа больших данных**

УТВЕРЖДАЮ

Проректор по учебной
и методической работе

_____ Е.А. Каменева

24.05.2024 г.

Макрушин С.В., Блохин Н.В.

Технологии обработки и анализа больших данных

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки:

38.03.01 - Экономика,

ОП «Бизнес-анализ, налоги и аудит»,

Профили: «Аудит и внутренний контроль», «Учёт, анализ и аудит»

*Рекомендовано Ученым советом
Факультета информационных технологий и анализа больших данных
(протокол № 44 от 21.05.2024 г.)*

*Одобрено советом Кафедры анализа данных и машинного обучения
(протокол № 01 от 06.05.2024 г.)*

Москва 2024

Содержание

1. Наименование дисциплины.....	2
2.Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине	2
3. Место дисциплины в структуре образовательной программы	3
4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся	3
5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий	4
5.1. Содержание дисциплины	4
5.2. Учебно-тематический план.....	8
5.3. Содержание семинаров, практических занятий	10
6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине.....	12
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	12
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю	13
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине.....	16
8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	21
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.....	22
10. Методические указания для обучающихся по освоению дисциплины .	24
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем	24
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.....	25

1. Наименование дисциплины

«Технологии обработки и анализа 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
Профиль: «Аудит и внутренний контроль»			
ПКП-5	Способность проводить мероприятия по внутреннему контролю, формированию информационной базы объекта внутреннего контроля, ее анализу	1.Проводит мероприятия по внутреннему контролю.	Знать: основные принципы внутреннего контроля, а также методы анализа данных для обнаружения аномалий и ошибок Уметь: проводить мероприятия по внутреннему контролю, включая анализ данных и выявление потенциальных рисков и проблем
		2.Формирует и анализирует информационную базу объектов внутреннего контроля.	Знать: методы формирования информационной базы объектов внутреннего контроля, включая сбор, хранение и организацию данных о контрольных мероприятиях и их результатов Уметь: анализировать информационную базу объектов внутреннего контроля с целью выявления трендов, аномалий и потенциальных рисков
Профиль: «Учет, анализ и аудит»			
ПКП-5	Способность к использованию специальных программных продуктов, применяемых для выполнения бухгалтерско-аналитических и контрольных функций в экономическом субъекте	1. Использует специальные программные продукты для выполнения бухгалтерско-аналитических и контрольных функций в экономическом субъекте.	Знать: специальные инструменты для обработки больших данных. Уметь: использовать специальные инструменты для обработки больших данных для выполнения бухгалтерско-аналитических функций.

		2. Демонстрирует владение специальными программными продуктами, применяемыми для выполнения бухгалтерско-аналитических и контрольных функций в экономическом субъекте.	Знать: специальные программные продукты для обработки больших объемов бухгалтерской информации. Уметь: использовать специальные программные продукты для обработки больших объемов бухгалтерской информации.
--	--	--	---

3. Место дисциплины в структуре образовательной программы

Дисциплина «Технологии обработки и анализа больших данных» является дисциплиной Цикла профиля (элективный) по направлению подготовки 38.03.01 – Экономика, ОП «Бизнес-анализ, налоги и аудит»,
Профили: «Аудит и внутренний контроль», «Учёт, анализ и аудит».

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

Профиль: «Аудит и внутренний контроль»

очная форма обучения

Вид учебной работы по дисциплине	Всего (в з.е. и часах)	Семестр 6 (в часах)
Общая трудоёмкость дисциплины	3/108	108
Контактная работа- Аудиторные занятия	34	34
Лекции	16	16
Семинары, практические занятия	18	18
Самостоятельная работа	74	74
Вид текущего контроля	Зачет	Зачет
Вид промежуточной аттестации	Контрольная работа	Контрольная работа

Профиль: «Учет, анализ и аудит»

очно-заочная форма обучения (ИОО)

Вид учебной работы по дисциплине	Всего (в з.е. и часах)	Семестр 7 (в часах)
Общая трудоёмкость дисциплины	3/108	108
Контактная работа-Аудиторные занятия	16	16
Лекции	8	8
Семинары, практические занятия	8	8
Самостоятельная работа	92	92
Вид текущего контроля	Зачет	Зачет
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. Объедине-

ние 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. 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 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. Параллельная обработка данных, введение в Dask

В рамках темы рассматривается специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. 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`.

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`.

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.

5.2. Учебно-тематический план

Профиль: «Аудит и внутренний контроль»

очная форма обучения

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемо- сти
		Всего	*Контактная работа- Аудиторная работа			Самосто- ятельная работа	
			Общая, в т.ч.:	Лекции	Семинары, практические занятия		
1	Библиотека NumPy и Pandas	16	6	2	4	10	Участие в решении задач на практических занятиях. Обсуждения по результатам самостоятельной работы
2	Использование различных форматов файлов в задачах обработки данных.	14	4	2	2	10	
3	Взаимодействие с табличными данными в приложениях обработки данных.	14	4	2	2	10	
4	Визуализация данных	12	4	2	2	8	
5	Работа со строками в приложениях обработки данных	12	4	2	2	8	Участие в решении задач на практических занятиях. Обсуждения по результатам самостоятельной работы
6	Взаимодействие с базой данных в приложениях обработки данных.	12	4	2	2	8	
7	Профилирование процессов обработки данных, библиотека	14	4	2	2	10	

	Numba и векторизация в Numpy и Numba						
8	Параллельная обработка данных, введение в Dask	14	4	2	2	10	
	В целом по дисциплине	108	34	16	18	74	Согласно учебному плану: контрольная работа
	Итого в %		31	47	53	69	

Профиль: «Учет, анализ и аудит»

очно-заочная форма обучения (ИОО)

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы текущего контроля успеваемо- сти
		Всего	*Контактная работа- Аудиторная работа			Самостоя- тельная работа	
			Общая, в т.ч.:	Лекции	Семинары, практиче- ские занятия		
1	Библиотека NumPy и Pandas	17	2	1	1	15	Участие в решении за- дач на прак- тических занятиях. Обсужде- ния по ре- зультатам самостоя- тельной ра- боты
2	Использование различных форматов файлов в задачах обработки данных.	17	2	1	1	15	
3	Взаимодействие с табличными данными в приложениях обработки данных.	14	2	1	1	12	
4	Визуализация данных	14	2	1	1	12	
5	Работа со строками в приложениях обработки данных	14	2	1	1	12	

6	Взаимодействие с базой данных в приложениях обработки данных.	12	2	1	1	10	Участие в решении задач на практических занятиях. Обсуждения по результатам самостоятельной работы
7	Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba	10	2	1	1	8	
8	Параллельная обработка данных, введение в Dask	10	2	1	1	8	
	В целом по дисциплине	108	16	8	8	92	Согласно учебному плану: контрольная работа
	Итого в %		15	50	50	85	

* объем контактной работы в очно-заочной/заочной формах обучения и индивидуальных учебных планах определяется соответствующими учебными планами. Темы, реализуемые в виде контактной работы, определяются преподавателем самостоятельно, исходя из уровня их сложности.

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	<ul style="list-style-type: none"> • 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. 	Интерактивная форма, работа на компьютере

	<ul style="list-style-type: none"> • 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]	
Использование различных форматов файлов в задачах обработки данных	<ul style="list-style-type: none"> • 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], 8[2], 9[3], 9[4]	Интерактивная форма, работа на компьютере
Взаимодействие с табличными данными в приложениях обработки данных.	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Основы работы с регулярными выражениями: базовый синтаксис, примеры. • Модуль re в Python. 8[1], 8[2], 9[4]	Интерактивная форма, работа на компьютере
Взаимодействие с базой данных приложениях обработки данных	<ul style="list-style-type: none"> • Взаимодействие из Python с базой данных с помощью API SQLite. 8[1], 8[2]	Интерактивная форма, работа на компьютере
Профилирование процессов обработки данных, библиотека Numba.	<ul style="list-style-type: none"> • 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
Параллельная обработка данных, введение в Dask	<ul style="list-style-type: none"> • специфика современного аппаратного обеспечения для обработки больших данных и проблема масштабируемости параллельных вычислений. • Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения. 	Интерактивная форма, работа на компьютере

	<ul style="list-style-type: none"> • Проблема Global Interpreter Lock в Python и способы обхода ее ограничений. • Модуль Python multiprocessing – назначение и основные возможности, API multiprocessing.Pool. • 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 <p>8[1], 8[2], 9[5], 9[7], 9[8], 9[10], 9[11], 9[15]</p>	
--	--	--

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	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • 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.	<ul style="list-style-type: none"> • Продвинутое взаимодействие с Excel из Python с помощью библиотеки XLWings. 	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.

Визуализация данных	<ul style="list-style-type: none"> • 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	<ul style="list-style-type: none"> • Использования хэширования при работе со строками. • Строки в библиотеке numpy. 	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Взаимодействие с базой данных в приложениях обработки данных	<ul style="list-style-type: none"> • Базовые возможности работы с транзакциями с помощью API SQLite. 	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Профилирование процессов обработки данных, библиотека Numba.	<ul style="list-style-type: none"> • 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.
Параллельная обработка данных, введение в Dask	<ul style="list-style-type: none"> • Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. • Специфика различия между потоками и процессами. • Организация вычислений с помощью Map / Filter / Reduce: общий принцип и специфика параллельной реализации обработки данных с помощью Dask.Bag. • Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение. 	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. Подходы к декомпозиции крупных вычислительных задач на подзадачи для параллельного исполнения

4. Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем
5. Различия между потоками и процессами, различие между различными планировщиками в Dask
6. Граф зависимостей задач – суть структуры данных, ее построение и использование в Dask
7. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач
8. Dask.Array – структура данных, специфика реализации и применения, процедура создания
9. Dask.Array – поддерживаемые операции и отличия от NumPy ndarray
10. Dask.DataFrame - структура данных, специфика реализации и применения, процедура создания Dask.DataFrame
11. Ограничения использования Dask.DataFrame и операции мэппинга в Dask.DataFrame
12. Поддержка Dask.DataFrame операций работающих со скользящим окном
13. Совместное использование промежуточных результатов в Dask: принцип работы и примеры использования
14. Dask.Bag - структура данных, специфика реализации и применения, процедура создания DaskBag
15. Организация вычислений с помощью Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных в Dask.Bag
16. 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).

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях Кафедры 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
Профиль: «Аудит и внутренний контроль»			
ПКП-5 Способность проводить мероприятия по внутреннему контролю, формированию информационной базы объекта внутреннего контроля, ее анализу	1.Проводит мероприятия по внутреннему контролю.	Знать: основные принципы внутреннего контроля, а также методы анализа данных для обнаружения аномалий и ошибок Уметь: проводить мероприятия по внутреннему контролю, включая анализ данных и выявление потенциальных рисков и проблем	Загрузите данные о выполненных заказах из базы данных, содержащей информацию о времени обработки заказа, количестве заказанных товаров и стоимости доставки. Рассчитайте среднее время обработки заказа, оцените процент заказов, выполненных с задержкой и проведите анализ стоимости доставки, выявив заказы с аномально высокими или низкими стоимостями доставки.
	2.Формирует и анализирует информационную базу объектов внутреннего контроля.	Знать: методы формирования информационной базы объектов внутреннего контроля, включая сбор, хранение и организацию данных о контрольных мероприятиях и их результатах Уметь: анализировать информационную базу объектов внутреннего контроля с целью выявления трендов, аномалий и потенциальных рисков	Разработайте программу на Python для формирования базы данных с информацией о контрольных мероприятиях и их результатах на основе файлов формата JSON и XML.

Профиль: «Учет, анализ и аудит»			
ПКП-5 Способность к использованию специальных программных продуктов, применяемых для выполнения бухгалтерско-аналитических и контрольных функций в экономическом субъекте	1.Использует специальные программные продукты для выполнения бухгалтерско-аналитических и контрольных функций в экономическом субъекте.	Знать: специальные инструменты для обработки больших данных. Уметь: использовать специальные инструменты для обработки больших данных для выполнения бухгалтерско-аналитических функций.	Предложите технологический стек для анализа первичной отчетности объемом 10Мб, 100Мб, 1Гб. Реализуйте алгоритм агрегации показателей для набора данных объемом 100 Мб.
	2.Демонстрирует владение специальными программными продуктами, применяемыми для выполнения бухгалтерско-аналитических и контрольных функций в экономическом субъекте.	Знать: специальные программные продукты для обработки больших объемов бухгалтерской информации. Уметь: использовать специальные программные продукты для обработки больших объемов бухгалтерской информации.	Предложите технологический стек для многокритериальной группировки данных первичной отчетности объемом 100Мб, Реализуйте алгоритм многокритериальной группировки данных первичной отчетности объемом 100 Мб.

Примерные вопросы для подготовки к зачету

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 взаимодействия 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 токенизация текста на естественном языке, стемминг и лемматизация, примеры на Python
28. Различия между потоками и процессами, различие между различными планировщиками в Dask
29. Граф зависимостей задач – суть структуры данных, ее построение и использование в Dask
30. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач
31. Dask.Array – структура данных, специфика реализации и применения, процедура создания
32. Dask.Array – поддерживаемые операции и отличия от NumPy ndarray
33. Распараллеливание алгоритмов с помощью dask.delayed – принцип и примеры использования
34. Дополнительные параметры декоратора dask.delayed – назначение и примеры использования

- 35. Использование `dask.delayed` для объектов и операции над объектами `dask.delayed`, включая ограничения их использования
- 36. `Dask.DataFrame` - структура данных, специфика реализации и применения, процедура создания `Dask.DataFrame`
- 37. Ограничения использования `Dask.DataFrame` и операции мэппинга в `Dask.DataFrame`
- 38. Поддержка `Dask.DataFrame` операций работающих со скользящим окном
- 39. Совместное использование промежуточных результатов в `Dask`: принцип работы и примеры использования
- 40. `Dask.Bag` - структура данных, специфика реализации и применения, процедура создания `Dask.Bag`
- 41. Организация вычислений с помощью `Map / Filter / Reduce` : общий принцип и специфика параллельной реализации обработки данных в `Dask.Bag`
- 42. Понятие признака в анализе данных и типы признаков
- 43. Понятие разведочного анализа данных, основные задачи и типовые визуализации для решения этих задач

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. - URL: <https://znanium.com/catalog/product/1230215> (дата обращения: 23.04.2024). – 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: 23.04.2024).

– 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 ОН-ЛАЙН» <http://biblioclub.ru/>
4. Электронно-библиотечная система Znanium <http://www.znanium.com>
5. Электронно-библиотечная система издательства «ЮРАЙТ» <https://urait.ru/>
6. Электронно-библиотечная система издательства Проспект <http://ebs.prospekt.org/books>
7. Электронно-библиотечная система издательства Лань <https://e.lanbook.com/>
8. Деловая онлайн-библиотека Alpina Digital <http://lib.alpinadigital.ru/>
9. Электронная библиотека Издательского дома «Гребенников» <https://grebennikon.ru/>
10. Математические журналы: полнотекстовая коллекция Математического института им. В.А. Стеклова РАН <https://www.mathnet.ru/>
11. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>
12. Национальная электронная библиотека <http://нэб.рф/>
13. Ресурсы информационно-аналитического агентства по финансовым рынкам Cbonds.ru <https://cbonds.ru/>
14. СПАРК <https://spark-interfax.ru/>
15. CNKI. Academic Reference <https://ar.oversea.cnki.net/>
16. Электронные продукты издательства Elsevier <http://www.sciencedirect.com>
17. Emerald: Management eJournal Portfolio <https://www.emerald.com/insight/>
18. Реферативная база данных по математике MathSciNET <https://mathscinet.ams.org/mathscinet/>

19. Коллекция научных журналов Oxford University Press
<https://academic.oup.com/journals/>
20. Электронные коллекции книг и журналов издательства Springer:
<http://link.springer.com/>
21. Платформа STATISTA <https://www.statista.com/>
22. База данных научных журналов издательства Wiley
<https://onlinelibrary.wiley.com/>
23. Pyru 1.0.9 [Электронный ресурс]: сайт. – Режим доступа:
<https://pypi.python.org/pypi/pyru>
24. Python Data Analysis Library [Электронный ресурс]: сайт. – Режим доступа: <http://pandas.pydata.org/>
25. Python Documentation [Электронный ресурс]: сайт. – Режим доступа:
<http://python.org/doc/>
26. Python Standard Library [Электронный ресурс]: сайт. – Режим доступа:
<https://docs.python.org/2/library/>
27. Scikit-learn Machine Learning in Python [Электронный ресурс]: сайт. – Режим доступа: <http://scikit-learn.org>
28. Официальный сайт продукта <https://www.python.org/>
29. Каталог курсов Интернет Университета Информационных Технологий
<http://www.intuit.ru/>
30. The Python Tutorial // <https://docs.python.org/3/tutorial/index.html>
31. NumPy User Guide // <http://docs.scipy.org/doc/numpy/user/index.html>
32. Pandas User Guide <http://pandas.pydata.org/pandas-docs/stable/>
33. Dask User Guide <https://docs.dask.org/en/latest/>
34. Matplotlib User Guide // <https://matplotlib.org/stable/users/index.html>
35. 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, пометая 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.