

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

Ярославский филиал Финуниверситета

Кафедра «Экономика и финансы»

СОГЛАСОВАНО

Генеральный директор
ПАО «Автодизель»
(Ярославский моторный завод)



А.А. Матюшин

УТВЕРЖДАЮ

Директор
Ярославского филиала
Финуниверситета



В.А. Кваша

«09» июня 2026 г.

«16» июня 2026 г.

Автор: О.Ю. Манина

Технологии обработки больших данных

Рабочая программа дисциплины

для студентов, обучающихся по направлению (ям) подготовки
38.03.01 «Экономика»,
профиль «Бизнес-архитектура и аналитика»
(очная форма обучения)

*Рекомендовано Ученым советом Ярославского филиала Финуниверситета
(протокол № 6 от 16.06.2026)*

*Одобрено кафедрой «Экономика и финансы»
(протокол № 6 от 04.06.2026)*

Ярославль 2026

СОДЕРЖАНИЕ

1. Наименование дисциплины.....	3
2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине.	3
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 (в семестре, в сессию).....	5
5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий.....	5
5.1. Содержание дисциплины	5
5.2. Учебно-тематический план	8
5.3. Содержание семинаров, практических занятий	9
6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине.....	11
6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы.....	11
6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю	13
7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине	17
7.1 Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы.....	17
7.2. Типовые контрольные задания или иные материалы, необходимые для оценки индикаторов достижения компетенций, умений и знаний	17
7.3. Соответствующие приказы, распоряжения ректората о контроле уровня освоения дисциплин и сформированности компетенций студентов.....	17
8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	22
9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины.....	22
10. Методические указания для обучающихся по освоению дисциплины	22
11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем.....	23
12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.....	24

1. Наименование дисциплины

Дисциплина Б.1.1.2.10 «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 формирование следующих компетенций.

Таблица 1 – Структура планируемых результатов обучения по дисциплине

Код компетенции	Наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
УК-4	Способность использовать прикладное программное обеспечение при решении профессиональных задач	1. Использует основные методы и средства получения, представления, хранения и обработки данных. 2. Демонстрирует владение профессиональными пакетами прикладных программ. 3. Выбирает необходимое прикладное программное обеспечение в зависимости от решаемой задачи. 4. Использует прикладное программное обеспечение	1.Знать: Основные методы и средства получения, представления, хранения и обработки данных, включая форматы файлов и библиотеки Python для работы с данными. Уметь: Применять библиотеки для работы с массивами данных, загружать, сохранять и обрабатывать данные из различных источников, применять различные методы обработки данных. 2.Знать: Профессиональные пакеты прикладных программ, используемые в соответствующей области, и их возможности для решения задач Уметь: Применять профессиональные пакеты прикладных программ для решения конкретных профессиональных задач 3.Знать: Основные характеристики различных прикладных программ и их область применения в контексте решаемых задач. Уметь: анализировать поставленную задачу и выбрать наиболее подходящее прикладное программное обеспечение для ее решения. 4.Знать: Различные прикладные программные решения, их функциональные возможности и области применения.

		для решения конкретных прикладных задач.	Уметь: Применять конкретное прикладное программное обеспечение для решения специфических задач, учитывая требования и особенности задачи.
УК-10	Способность осуществлять поиск, критически анализировать, обобщать и систематизировать информацию, использовать системный подход для решения поставленных задач	1. Четко описывает состав и структуру требуемых данных и информации, грамотно реализует процессы их сбора, обработки и интерпретации 2. Обосновывает сущность происходящего, выявляет закономерности, понимает природу вариабельности 3. Формулирует признак классификации, выделяет соответствующие ему группы однородных «объектов», идентифицирует общие свойства элементов этих групп, оценивает полноту результатов классификации, показывает прикладное назначение классификационных групп. 4. Грамотно, логично, аргументировано формирует собственные суждения и оценки.	1.Знать: Структуру и состав необходимых данных и информации для решения конкретных задач, а также методы их сбора, обработки и интерпретации. Уметь: Описывать требуемую структуру данных и информации, проводить систематический поиск необходимой информации, анализировать её критически и систематизировать для использования в решении задач. 2.Знать: Основные концепции и теории, связанные с анализом данных и системным подходом к решению задач. Уметь: Обосновывать сущность происходящего, выявлять закономерности и понимать природу вариабельности данных при анализе информации. 3.Знать: Основные принципы классификации данных, методы формулировки признаков классификации и принципы выделения групп однородных объектов. Уметь: Формулировать признаки классификации на основе анализа данных, выделять соответствующие группы однородных объектов и идентифицировать их общие свойства. 4.Знать: Основные принципы логического аргументирования на основе интерпретации информации.

		Отличает факты от мнений, интерпретаций, оценок и т.д. в рассуждениях других участников деятельности. 5. Аргументированно и логично представляет свою точку зрения посредством и на основе системного описания.	Уметь: Формировать собственные суждения и оценки на основе критического анализа информации, грамотно и логично аргументировать свои выводы. 5. Знать: Методы системного описания информации. Уметь: представлять свою точку зрения, используя соответствующие факты, данные и аналитические выводы.
--	--	--	--

3. Место дисциплины в структуре образовательной программы

Дисциплина «Технологии обработки больших данных» входит в цикл математики и информатики модуля общепрофессиональных дисциплин обязательной части образовательной программы, по направлению 38.03.01 Экономика, профиль «Бизнес-архитектура и аналитика», очная форма обучения.

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

Таблица 2 – Объем дисциплины в зачетных единицах и академических часах

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 7 в (часах)
Общая трудоёмкость дисциплины	5 з.е./180 ч.	5 з.е./180 ч.
<i>Контактная работа – Аудиторные занятия</i>	68	68
<i>Лекции</i>	24	24
<i>Семинары, практические занятия</i>	44	44
<i>Самостоятельная работа</i>	112	112
Вид текущего контроля	Контр.раб.	Контр.раб.
Вид промежуточной аттестации	Экзамен	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 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 цвето-

вые 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

Таблица 3 – Содержание учебно-тематического плана изучения дисциплины

№ п/п	Наименование тем (разделов) дисциплины	Трудоемкость в часах					Формы теку- щего контроля успеваемости
		Всег о	Контактная работа* Аудиторная работа			Самосто- ятельная работа	
			Об- щая, в т.ч.:	Лек- ции	Семинары, практические занятия		
1.	Библиотека NumPy и Pandas	23	9	3	6	14	Участие в реше- нии задач на практических за- нятиях. Обсуж- дения по резуль- татам самостоя- тельной работы
2.	Использование различных фор- матов файлов в задачах обра- ботки данных.	23	9	3	6	14	
3.	Взаимодействие с табличными данными в при- ложениях обра- ботки данных.	23	9	3	6	14	
4.	Визуализация данных	21	7	3	4	14	

5.	Работа со строками в приложениях обработки данных	23	9	3	6	14	
6.	Взаимодействие с базой данных в приложениях обработки данных.	23	9	3	6	14	
7.	Профилирование процессов обработки данных, библиотека Numba и векторизация в Numpy и Numba	19	5	3	2	14	
8.	Параллельная обработка данных, введение в Dask	25	11	3	8	14	
	В целом по дисциплине	180	68	24	44	112	
	Итого в %		100		65		

5.3. Содержание семинаров, практических занятий

Таблица 4 – Темы дисциплины, формы проведения и содержание семинаров и 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. 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.	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
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
Профилирование процессов обработки данных, библиотека 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

Параллельная обработка данных, введение в Dask	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. 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 8[1], 8[2], 9[5], 9[7], 9[8], 9[10], 9[11],9[15]	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

К внеаудиторным формам самостоятельной работы студентов относятся:

- подготовка к семинарскому занятию;
- подготовка к решению и проверке задач;
- подготовка контрольной работы;
- подготовка к экзамену.

На семинарских занятиях все студенты должны принимать активное участие в обсуждении изучаемых вопросов и уметь демонстрировать знание практического материала. При выступлении, студентам необходимо аргументировано излагать свою позицию, подкреплять ее конкретными данными, уметь обобщать, аргументировать и систематизировать статистические данные.

На семинарских занятиях преподаватель проверяет выполнение самостоятельных заданий и качество усвоения знаний.

Таблица 5 – Темы дисциплины, изучаемые вопросы и формы внеаудиторной самостоятельной работы

Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы
--	--	--

Библиотека NumPy и Pandas	• Принцип распространения значений при выполнении операций в NumPy: общий алгоритм и примеры. • Маскирование и прихотливое индексирование в NumPy. • Операция GroupBy в Pandas DataFrame и реализация в ней подхода «разбиение, применение и объединение».	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Использование различных форматов файлов в задачах обработки данных	• Формат файлов NPY, представление данных в этом формате и взаимодействие с ним в Python. • Формат файлов HDF, представление данных в этом формате и взаимодействие с ним в Python.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Взаимодействие с табличными данными в приложениях обработки данных.	Продвинутое взаимодействие с Excel из Python с помощью библиотеки XLWings.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Визуализация данных	Построение трехмерных графиков Продвинутое взаимодействие с цветовыми картами	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Работа со строками в приложениях обработки данных	• Использование 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	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.	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	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. Машинное обучение при работе с Dask	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.

Оценка знаний студентов осуществляется в баллах с учетом оценки работы в семестре (написание реферата, аудиторных самостоятельных работ и домашних заданий, решение задач и участие в обсуждениях на практических занятиях и др.), оценки итоговых знаний (по результатам зачета) и в соответствии с критериями Финансового университета реализуется следующим образом:

№п/п	Вид отчетности	Баллы
1.	Работа в семестре	40
	В том числе расчетно-аналитическая работа	10
2.	Экзамен	60
	Итого:	100

Таблица 6 - Формы текущего контроля успеваемости и их балльная оценка

Критерий оценки	Баллы	Методика оценки
Посещение аудиторных занятий	7	Математический расчет $Оц = 7 : Кз \times Пз$ Где Оц – сумма баллов; Кз – количество аудиторных занятий по дисциплине; Пз – количество занятий, посещаемых студентом
Активность при проведении аудиторных занятий	13	Доклад, презентация – 2 балла Участие в обсуждении – 2 балла Решение расчетных задач – 2 балла
Активность при проведении вне-аудиторной работы	10	Подготовка материалов и выступление на студенческих научных конференциях, в

		печатных изданиях, в том числе и Университета - 10 баллов
Самостоятельная работа студента под руководством преподавателя	10	Выполнение письменной работы – 10 баллов

С целью стимулирования систематической подготовки студентов к практическим и семинарским занятиям по дисциплине в течение семестра вводится комплексный подход к оценке, получаемой студентами по итогам изучения дисциплины. На основании положения о системе оценки знаний студентов в Финансовом университете действует 100-балльная система оценки знаний. Это означает, что оценка, получаемая по итогам письменной экзаменационной работы, состоит из двух частей: текущего контроля студентов в семестре — максимальная оценка 40 баллов и результатов промежуточной аттестации — максимальная оценка 60 баллов.

Текущий контроль осуществляется в ходе учебного процесса и консультирования студентов, по результатам выполнения самостоятельных работ. Основными формами текущего контроля знаний являются:

- решение задач, тестов и их обсуждение с точки зрения умения формулировать выводы, вносить рекомендации и принимать адекватные управленческие решения;
- выполнение контрольных заданий и обсуждение результатов;
- обсуждения законодательных, правовых и нормативных актов.

Вторая составляющая оценивает устный ответ на экзамене и не может превышать 60 баллов.

40 баллов, полученные студентом в течение семестра, должны означать самую высокую характеристику его работы. Такой балл получают студенты, которые на практических занятиях систематически показывают высокие результаты при опросах, проявляют активность при обсуждении изучаемых проблем, в полном объеме выполняют учебную программу, не имеют пропусков.

Максимальная оценка ответа на экзамене оценивается в 60 баллов. Каждый вопрос в билете имеет свою долю в общей оценке в зависимости от сложности. Например, за первый теоретический вопрос максимально оценивается 15 баллов, два следующих теоретических вопроса оцениваются по 10 баллов, задача — максимально 15 баллов и пять тестов — по 2 балла каждый правильно решенный тест.

Балльная система аттестации:

- студент, получивший от 50 до 100 баллов, считается аттестованным, получивший от 0 до 49 баллов - не аттестованным;
- студент, получивший от 50 до 69 баллов, считается аттестованным на «удовлетворительно»;
- студент, получивший от 70 до 85 баллов, считается аттестованным на «хорошо»;
- студент, получивший от 86 до 100 баллов, считается аттестованным на «отлично».

Таким образом, если студент по итогам работы в семестре набирает 21 балл, а по итогам письменной работы — 60 баллов, то общая сумма 81 балл соответствует окончательной оценке 4 «хорошо».

Если студент подошел к экзамену с оценкой 0 баллов, то при безупречном качестве письменной экзаменационной работы он может получить итоговую оценку по данной дисциплине только 60 баллов, что соответствует оценке «удовлетворительно».

О данном подходе к оценке знаний студентов преподаватель информирует студентов на первом семинарском (практическом) занятии. На последнем семинарском занятии студентам сообщается оценка, которую они получают по итогам работы в семестре.

Студенты могут улучшить свою оценку по итогам работы в семестре за счет отработки пропущенных занятий или полученных неудовлетворительных оценок на семинарах. Вместе с тем, все отработки могут быть осуществлены в консультационные дни в период не более двух консультаций с момента получения оценки «неудовлетворительно» или пропуска занятий. Только в случае болезни студент может отрабатывать пропуски в более поздние сроки.

Отработка пропусков, имевших место по причине работы студентов во время занятий, не допускается.

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

7.1 Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы

Перечень компетенций, формируемых в процессе освоения дисциплины, содержится в разделе 2 «Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине».

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

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 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. Расстояние Левенштейна: 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

- ## Пример экзаменационного билета

Кафедра	<u>«Экономика и финансы»</u>		
Дисциплина	<u>«Технологии обработки больших данных»</u>		
Филиал	<u>Ярославский</u>	Форма обучения	<u>Очная</u>
Семестр	7	Направление	38.03.01 «Экономика»
Профиль	«Бизнес-архитектура и аналитика»		

№	Вопрос билета	Макс. балл
1	Большие данные – определение и причины возникновения задач обработки больших данных	20
2	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.	20

	2. Которая по имени плейлиста возвращает количество треков в нем и их суммарную продолжительность.	
3	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).	20

Подготовил

Манина О.Ю.

Утверждаю:

Заведующий кафедрой
«Экономика и финансы»

Сироткин С.А.

Дата _____ 20__ г.

Примеры оценочных средств для проверки компетенций, формируемых дисциплиной

Наименование компетенции	Типовые контрольные задания
УК-4 Способность использовать прикладное программное обеспечение при решении профессиональных задач	1.Использует основные методы и средства получения, представления, хранения и обработки данных. Напишите программу на Python, которая загружает данные из CSV-файла, проводит анализ их средствами библиотеки Pandas, а затем сохраняет результаты анализа в новый файл 2. Демонстрирует владение профессиональными пакетами прикладных программ. Напишите программу на Python, которая загружает данные о цене акций из CSV-файла для создания временного ряда цен акций и визуализирует его на графике при помощи Matplotlib. 3.Выбирает необходимое прикладное программное обеспечение в зависимости от решаемой задачи. Напишите программу на Python, которая загружает данные о ценах акций из базы данных, использует подходящие инструменты анализа данных и подготавливает отчет с графиками и аналитическими выводами о динамике ценных бумаг на рынке. 4.Использует прикладное программное обеспечение для решения конкретных прикладных задач. Напишите программу на Python, которая загружает данные о рекламных расходах и результате кампании из базы данных, использует библиотеку Pandas для анализа данных и библиотеку Matplotlib для визуализации ключевых метрик эффективности кампании (например, ROI, конверсионные показатели).

УК-10 Способность осуществлять поиск, критически анализировать, обобщать и систематизировать информацию, использовать системный подход для решения поставленных задач	1. Четко описывает состав и структуру требуемых данных и информации, грамотно реализует процессы их сбора, обработки и интерпретации Подготовить план действий, описывающий процессы сбора, обработки и интерпретации необходимой информации о текущем состоянии рынка и конкурентной среды для выработки стратегии развития бизнеса на следующий год. План должен включать источники данных (например, открытые базы данных, отчеты аналитических агентств, внутренние данные компании), методы анализа и формат представления результатов. 2. Обосновывает сущность происходящего, выявляет закономерности, понимает природу вариабельности Рассмотрите набор данных о финансовых показателях компаний из разных секторов экономики за последние пять лет. Опишите основные тенденции и закономерности, выявленные в данных. 3. Формулирует признак классификации, выделяет соответствующие ему группы однородных «объектов», идентифицирует общие свойства элементов этих групп, оценивает полноту результатов классификации, показывает прикладное назначение классификационных групп. Сформулируйте признаки классификации клиентов интернет-магазина (например, по частоте покупок, сумме трат, предпочтениям в товарах). Выделите соответствующие группы однородных клиентов и идентифицируйте их общие характеристики. Для решения используйте Pandas. 4. Грамотно, логично, аргументировано формирует собственные суждения и оценки. Отличает факты от мнений, интерпретаций, оценок и т.д. в рассуждениях других участников деятельности. Рассмотрите отчет о рыночной конкуренции в отрасли. Оцените представленную информацию. Проанализируйте данные при помощи Python и сформулируйте оценку текущего состояния конкуренции в отрасли. 5. Аргументированно и логично представляет свою точку зрения посредством и на основе системного описания. Используя библиотеки Pandas и Matplotlib, проанализируйте данные о рынке труда в отрасли, включая статистику безработицы, занятости и средних зарплат за последние несколько лет. На основе полученных данных логически и аргументированно опишите свою точку зрения о текущей ситуации на рынке труда.
ПКН-5 Способность составлять и анализировать финансовую, бухгалтерскую, статистическую отчетность и использовать результаты анализа для принятия управленческих решений	1. Применяет положения международных и национальных стандартов для составления и подтверждения достоверности отчетности организации. Проанализировать большой объем финансовых данных организации. Проверить достоверность данных и выявить закономерности и тенденции, которые могут влиять на финансовое положение организации. Для анализа данных используйте Dask. 2. Использует результаты анализа финансовой, бухгалтерской, статистической отчетности при составлении финансовых планов, отборе инвестиционных проектов и принятии оперативных решений на макро-, мезо- и микроуровнях. Используя библиотеку Pandas, проведите анализ компании за последний финансовый год. Оцените ключевые показатели эффективности и рисков. На основе анализа сформируйте финансовый план на следующий год с использованием библиотеки Matplotlib для визуализации ключевых метрик.

7.3. Соответствующие приказы, распоряжения ректората о контроле уровня освоения дисциплин и сформированности компетенций студентов.

Приказ от 01 октября 2024 года №2187/о «Об утверждении Положения о проведении текущего контроля успеваемости и промежуточной аттестации студентов, обучающихся по образовательным программам высшего образования в Финансовом университете», размещенный по ссылке: https://www.fa.ru/upload/iblock/845/yqw4vi0gi21glvu9775bfm4rz6p19b0z/Prikaz_-_2187_o-ot-01.10.2024.pdf

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: 24.04.2024). – 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: 24.04.2024). – Текст : электронный.

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

1. Информационно-образовательный портал Финансового университета при Правительстве Российской Федерации. <http://portal.ufrf.ru/>.
2. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/>
6. Электронно-библиотечная система BOOK.RU <http://www.book.ru>
7. Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» <http://biblioclub.ru/>
8. Электронно-библиотечная система Znanium <http://www.znanium.com>
9. Образовательная платформа Юрайт <https://urait.ru/>
10. Электронно-библиотечная система издательства Проспект <http://ebs.prospekt.org/books>
11. Электронно-библиотечная система издательства «Лань» <https://e.lanbook.com/>
12. Электронная библиотека Издательского дома «Гребенников» <https://grebennikon.ru/>

13. Деловая онлайн-библиотека Alpina Digital <http://lib.alpinadigital.ru/>
14. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>
15. Национальная электронная библиотека <http://нэб.рф/>
16. Ресурсы информационно-аналитического агентства по финансовым рынкам Cbonds.ru <https://cbonds.ru/>
17. СПАРК <https://spark-interfax.ru/>
18. Academic Reference <http://ar.cnki.net/ACADREF>
19. Электронная коллекция книг издательства Springer: Springer eBooks [http://link.springer.com/Международного института бизнес-анализа](http://link.springer.com/Международного_института_бизнес-анализа). URL: <http://www.iiba.ru>

10. Методические указания для обучающихся по освоению дисциплины

Наименование методических материалов для обучающихся	Год утверждения	Местонахождение материала (ссылка на сайт филиала)
Манина О.Ю.. Методические рекомендации по выполнению контрольной работы по дисциплине «Технологии обработки больших данных» для студентов, обучающихся по направлению подготовки 38.03.01 «Экономика», профиль Бизнес-архитектура и аналитика	2026	https://cloud.mail.ru/public/CnoV/cYBRvyAHR
Приказ Финуниверситета от 14.05.2021 г. № 1040/о «Об утверждении методических рекомендаций по планированию и организации внеаудиторной самостоятельной работы студентов по образовательным программам бакалавриата и магистратуры Финансовом университете»		https://www.fa.ru/upload/iblock/dc2/a5qpi6qb0k0l6jz01c4g9y1cv5x35ey3/Prikaz-1040_o-ot-11.05.2021.PDF

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. Linux, Libre Office.
2. Антивирус Kaspersky
- R и R-Studio
- Python

11.2. Современные профессиональные базы данных и информационные справочные системы:

11.3. Сертифицированные программные и аппаратные средства защиты информации: не предусмотрены.

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

Материально-техническая база, которой располагает Ярославский филиал Финуниверситета: аудиторный фонд, компьютерные классы и др.

Лекционные и практические занятия проводятся в мультимедийных аудиториях филиала, а также в оборудованных компьютерами и мультимедийным оборудованием классах.

Профессорско-педагогический состав знакомится с психолого-физиологическими особенностями обучающихся инвалидов и лиц с ограниченными возможностями здоровья. При необходимости осуществляется дополнительная поддержка преподавания тьюторами, психологами, социальными работниками, прошедшими подготовку ассистентами.

В соответствии с методическими рекомендациями Минобрнауки России (утверждены 08.04.2014 № АК-44/05вн) в курсе предполагается использовать социально-активные и рефлексивные методы обучения, технологии социокультурной реабилитации с целью оказания помощи в установлении полноценных межличностных отношений с другими студентами, создании комфортного психологического климата в студенческой группе. Подбор и разработка учебных материалов производятся с учетом предоставления материала в различных формах: аудиальной, визуальной, с использованием специальных технических средств и информационных систем.

Освоение дисциплины лицами с ОВЗ осуществляется с использованием средств обучения общего и специального назначения (персонального и коллективного использования). Материально-техническое обеспечение предусматривает приспособление аудиторий к needs лиц с ОВЗ.

Форма проведения аттестации для студентов-инвалидов устанавливается с учетом индивидуальных психофизических особенностей. Для студентов с ОВЗ предусматривается доступная форма предоставления заданий оценочных средств, а именно:

- в печатной или электронной форме (для лиц с нарушениями опорно-двигательного аппарата);
- в печатной форме или электронной форме с увеличенным шрифтом и контрастностью (для лиц с нарушениями слуха, речи, зрения);
- методом чтения ассистентом задания вслух (для лиц с нарушениями зрения).

Студентам с инвалидностью увеличивается время на подготовку ответов на контрольные вопросы. Для таких студентов предусматривается доступная форма предоставления ответов на задания, а именно:

- письменно на бумаге или набором ответов на компьютере (для лиц с нарушениями слуха, речи);
- выбором ответа из возможных вариантов с использованием услуг ассистента (для лиц с нарушениями опорно-двигательного аппарата);

– устно (для лиц с нарушениями зрения, опорно-двигательного аппарата).

При необходимости для обучающихся с инвалидностью процедура оценивания результатов обучения может проводиться в несколько этапов.