

Федеральное государственное образовательное бюджетное учреждение
высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)**

**Кафедра информационных технологий
Факультет информационных технологий и анализа больших данных**

Документ подписан усиленной неквалифицированной электронной подписью
Организация: Финансовый университет при Правительстве РФ
Утверждено: Проректор по учебной и методической работе Е.А. Каменева
Сертификат: ke7L25wwwxzjqckGLbZN5+g4cY0S+2kof
Дата: 26.11.2025 г.

Л.Н. Чернышов

Технологии обработки больших данных

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки:

09.03.03 - Прикладная информатика,

Образовательная программа «Прикладные информационные системы в экономике
и финансах»

Профиль:

«Инженерия данных»

Рекомендовано

*Факультет информационных технологий и анализа больших данных
(протокол № 03 от 16.12.2025 г.)*

Одобрено

*Кафедра информационных технологий
(протокол № 12 от 03.12.2025 г.)*

© Москва 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	4
4.	Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся	4
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	9
6.	Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине	10
6.1.	Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	10
6.2.	Перечень вопросов, заданий, тем для подготовки к текущему контролю	12
7.	Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине	13
8.	Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	19
9.	Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	20
10.	Методические указания для обучающихся по освоению дисциплины	21
11.	Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем	22
12.	Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине	23

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 технические 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» относится к «Профилю "Инженерия данных"».

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

Очная форма обучения

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 5 (в часах)
Общая трудоёмкость дисциплины	4/144	144
Контактная работа- Аудиторные занятия	50	50
Лекции	16	16
Семинары, практические занятия	34	34
Самостоятельная работа	94	94
Вид текущего контроля	Проектная работа	Проектная работа
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. Взаимодействие с базой данных в приложениях обработки данных

Взаимодействие из Python с базой данных на примере API SQLite. Базовые возможности работы с транзакциями. Парсинг web-страниц для извлечения данных в приложениях на Python с помощью библиотеки BeautifulSoup.

Тема 2. 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 3. Библиотека Dask

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

Рассматривается принцип и примеры использования распараллеливание алгоритмов с помощью dask.delayed. Рассматривается структура данных Dask.Array, специфика ее реализации и применения, процедура создания, поддерживаемые Dask.Array операции и ее отличия от NumPy ndarray. Рассматривается структура 5 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 4. 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. 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.

Тема 5. Apache Hadoop

В рамках данной темы рассматривается один из ключевых инструментов в области обработки больших данных - Hadoop. История и развитие Apache Hadoop. Хранение данных в распределенной среде (Hadoop Distributed File System). Hadoop MapReduce - модель программирования для обработки данных. YARN (Yet Another Resource Negotiator) - управление ресурсами в кластере. Компоненты экосистемы Hadoop: HBase, Hive, Spark. Процедура установки Hadoop на кластере и его основные компоненты. Обеспечение отказоустойчивости в Hadoop. Оптимизация производительности и масштабируемости.

Тема 6. Apache Spark

В рамках данной темы рассматриваются основы работы с Apache Spark, мощным фреймворком для обработки и анализа больших объемов данных, входящим в экосистему Apache Hadoop. Определение Apache Spark как фреймворка для распределенной обработки данных. История и эволюция Apache Spark в контексте технологий обработки данных. Основные компоненты Apache Spark. Распределенное хранение данных с использованием Resilient Distributed Datasets (RDD). Основы функционального программирования и операций над данными в Spark. Разработка приложений с использованием PySpark. Использование Spark Streaming для обработки потоковых данных. Введение в машинное обучение с использованием MLlib.

Взаимодействие с HDFS и Apache HBase. Интеграция с экосистемой Hadoop и инструментами для бизнес-аналитики.

5.2. Учебно-тематический план

№ п/п	Наименование тем(разделов) дисциплины	Трудоемкость в часах				
		Всего	Контактная работа - Аудиторная работа			Самостоя тельная работа
			Общая, в т.ч.:	Лекции	Семинары, практическ ие занятия	
1	Взаимодействие с базой данных в приложениях обработки данных	18	4	1	3	14
2	Параллельная обработка данных	18	5	2	3	13
3	Библиотека Dask	40	12	4	8	28
4	Обзор проблем обработки больших данных и вычисления общего назначения на GPU	12	3	1	2	9
5	Apache Hadoop	34	14	4	10	20
6	Apache Spark	22	12	4	8	10
	Итого	144	50	16	34	94

5.3. Содержание семинаров

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарах, практических занятиях	Формы проведения занятий
Взаимодействие с базой данных в приложениях обработки данных	Взаимодействие из Python с базой данных с помощью API SQLite.	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
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.	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.	Интерактивная форма, работа на компьютере
Apache Hadoop	Хранение данных в распределенной среде (Hadoop Distributed File System). Hadoop MapReduce. Процедура установки Hadoop на кластере.	Интерактивная форма, работа на компьютере
Apache Spark	Распределенное хранение данных с использованием Resilient Distributed Datasets (RDD). Основы функционального программирования и операций над данными в Spark. Использование Spark Streaming для обработки потоковых данных. Разработка приложений с использованием PySpark.	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
Взаимодействие с базой данных в приложениях обработки данных	Базовые возможности работы с транзакциями с помощью API SQLite.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Параллельная обработка данных	Модели параллельного программирования и их сочетаемость с архитектурами параллельных вычислительных систем. Специфика различия между потоками и процессами. Многопроцессорные архитектуры с общей и разделяемой памятью – специфика и сравнение.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Библиотека Dask	Организация вычислений с помощью Map / Filter / Reduce: общий принцип и специфика параллельной реализации обработки данных с помощью Dask. Bag. Организация вычислений с помощью API Dask Delayed.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Обзор проблем обработки больших данных и вычисления общего назначения на GPU	Специфика аппаратного обеспечения для решения задач обработки больших данных. Проблема выбора типичных средств обработки данных, адекватных различным объемам данных. Знакомство с библиотекой PyTorch. Понятие тензора в PyTorch. Базовые операции с тензорами в PyTorch.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.
Apache Hadoop	YARN (Yet Another Resource Negotiator) - управление ресурсами в кластере. Обеспечение отказоустойчивости в Hadoop. Оптимизация производительности и масштабируемости.	Обзор литературы и веб-источников. Самостоятельное освоение

		инструментов аналитической обработки. Решение задач.
Apache Spark	Введение в машинное обучение с использованием MLlib. Интеграция с экосистемой Hadoop и инструментами для бизнес-аналитики. Взаимодействие с HDFS и Apache HBase.	Обзор литературы и веб-источников. Самостоятельное освоение инструментов аналитической обработки. Решение задач.

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

Примерная тематика проектной работы

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 Кафедры информационных технологий 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

ПКП-2 Способность разрабатывать, согласовывать и управлять исполнением технического задания и технического проекта с использованием технологий больших данных

1) Работает со стандартами, в том числе адаптирует стандарты для специфических требований больших данных

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

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

Типовые контрольные задания

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.

2) Разрабатывает технические задания и технические проекты для технологий больших данных

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: технологию разработки технических заданий и технических проектов, в которых используются технологии больших данных

Уметь: разрабатывать технические задания и технические проекты, в которых используются технологии больших данных

Типовые контрольные задания

Описать технологию создания технического задания и технического проекта с учетом специфики больших данных.

3) Реализует управление рабочими проектами технологической инфраструктуры больших данных

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: современные принципы управления рабочими проектами, применяемыми к технологической инфраструктуре больших данных

Уметь: применять современные принципы управления рабочими проектами технологической инфраструктуры больших данных

Типовые контрольные задания

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.

Примеры практико-ориентированных заданий

Задание 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. 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 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 PySpark (корректным!), распараллелив процесс обработки данных (использование PySpark 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 PySpark, распараллелив процесс обработки данных (использование PySpark должно приводить к истинной параллельной обработке 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. Взаимодействие из Python с базой данных на примере API SQLite, базовые возможности работы с транзакциями
10. Модуль multiprocessing – назначение и основные возможности, API multiprocessing.Pool
11. Различия между потоками и процессами, различие между различными планировщиками в Dask
12. Граф зависимостей задач – суть структуры данных, ее построение и использование в Dask
13. Три ключевых структуры данных Dask: их специфика и принцип выбора структуры данных при решении задач
14. Dask.Array – структура данных, специфика реализации и применения, процедура создания
15. Dask.Array – поддерживаемые операции и отличия от NumPy ndarray
16. Распараллеливание алгоритмов с помощью dask.delayed – принцип и примеры использования
17. Дополнительные параметры декоратора dask.delayed – назначение и примеры использования
18. Использование dask.delayed для объектов и операции над объектами dask.delayed, включая ограничения их использования
19. Dask.DataFrame - структура данных, специфика реализации и применения, процедура создания Dask.DataFrame
20. Ограничения использования Dask.DataFrame и операции мэппинга в Dask.DataFrame
21. Поддержка Dask.DataFrame операций работающих со скользящим окном
22. Совместное использование промежуточных результатов в Dask: принцип работы и примеры использования

23. Dask.Bag - структура данных, специфика реализации и применения, процедура создания DaskBag
24. Организация вычислений с помощью Map / Filter / Reduce : общий принцип и специфика параллельной реализации обработки данных в Dask.Bag
25. API Dask.Bag – функции мэппинга, фильтрации и преобразования
26. API Dask.Bag – функции группировки и свертки
27. История и развитие Apache Hadoop, его роль в обработке больших данных
28. Компоненты Apache Hadoop и их предназначение
29. Архитектуру Hadoop Distributed File System (HDFS) и обеспечение отказоустойчивости данных в кластере
30. Основные компоненты Apache Spark и их взаимодействие для обработки данных
31. Resilient Distributed Datasets (RDD) в Apache Spark и операции можно выполнять над ними
32. Применение Apache Spark для обработки данных в реальном времени. Spark Streaming
33. Компоненты экосистемы Apache Spark для машинного обучения и их применение

Пример экзаменационного билета

Экзаменационный билет №

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. Кораблин, Ю. П. Структуры и алгоритмы обработки данных [Электронный ресурс] : учебно-методическое пособие / Кораблин Ю. П., Сыромятников В. П., Скворцова Л. А. Москва : РТУ МИРЭА, 2020 219 с. Книга из коллекции РТУ МИРЭА - Информатика <https://e.lanbook.com/book/163860>. [БИК ID: RU-LAN-BOOK-163860]
2. Железнов, М. М. Методы и технологии обработки больших данных [Электронный ресурс] : учебно-методическое пособие / Железнов М. М. Москва : МИСИ – МГСУ, 2020 46 с. Книга из коллекции МИСИ – МГСУ - Информатика <https://e.lanbook.com/book/145102> ISBN 978-5-7264-2193-3. [БИК ID: RU-LAN-BOOK-145102]

Дополнительная литература:

1. Атаян, А. М. Аналитика больших данных [Электронный ресурс] : учебно-методическое пособие / Атаян А. М., Вольфсон М. Б. Санкт-Петербург : СПбГУТ им. М.А. Бонч-Бруевича, 2023 69 с. Книга из коллекции СПбГУТ им. М.А. Бонч-Бруевича - Информатика <https://e.lanbook.com/book/426077>. [БИК ID: RU-LAN-BOOK-426077]
2. Пальмов, С. В. Основы сбора и обработки больших данных [Электронный ресурс] : учебное пособие / Пальмов С. В. Самара : ПГУТИ, 2023 285 с. Книга из коллекции ПГУТИ - Информатика <https://e.lanbook.com/book/411830>. [БИК ID: RU-LAN-BOOK-411830]
3. Новожилов, Олег Петрович Архитектура ЭВМ и вычислительных систем : учебник для вузов / О. П. Новожилов. 2-е изд., испр. и доп Электрон. дан. Москва : Юрайт, 2025 505 с (Высшее образование) URL: <https://urait.ru/bcode/568920> (дата обращения: 01.09.2025). Режим доступа: Электронно-библиотечная система Юрайт, для авториз. пользователей <https://urait.ru/bcode/568920> ISBN 978-5-534-20365-3 : 2399.00. [БИК ID: RU2fURAIT2f568920]
4. Гулаков, В. К. Структуры и алгоритмы обработки многомерных данных [Электронный ресурс] : монография / Гулаков В. К., Трубаков А. О., Трубаков Е. О. ; Гулаков В. К., Трубаков Е. О. 3-е изд., стер. Санкт-Петербург : Лань, 2024 356

с. Книга из коллекции Лань - Информатика <https://e.lanbook.com/book/388682> ISBN 978-5-507-49408-8. [БИК ID: RU-LAN-BOOK-388682]

5. Ланских, Ю. В. Введение в большие данные [Электронный ресурс] : учеб. пособие / Ланских Ю. В., Ланских В. Г., Родионов К. В. Киров : ВятГУ, 2023 172 с. Книга из коллекции ВятГУ - Информатика <https://e.lanbook.com/book/408566>. [БИК ID: RU-LAN-BOOK-408566]

6. Лебедев, А. С. Методы Big Data [Электронный ресурс] : учебно-методическое пособие / Лебедев А. С., Магомедов Ш. Г. Москва : РТУ МИРЭА, 2021 91 с. Книга из коллекции РТУ МИРЭА - Информатика <https://e.lanbook.com/book/182452>. [БИК ID: RU-LAN-BOOK-182452]

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

1. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/> (<http://library.fa.ru/files/elibfa.pdf>)
2. <https://www.planetaexcel.ru/>
3. Электронно-библиотечная система BOOK.RU <http://www.book.ru>

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.

Методические рекомендации по выполнению проектной работы:

Цель проектной работы - продемонстрировать владение основными навыками работы с большими данными, владение основными инструментальными средствами библиотек языка программирования Python, методами и приемами подготовительного и описательного анализа данных, средствами визуализации данных, умение делать выводы из проведенного анализа.

Последовательность выполнения работы:

- Выбрать набор данных для анализа в соответствии с выбранной темой проектной работы. Описать этот набор и решаемую задачу.
- Провести предварительный анализ и очистку данных. Этот этап включает в себя вывод информации о количественных характеристиках датасета, информацию об отсутствующих значениях, характеристиках и физическом смысле каждого атрибута данных, его значимости для предсказания целевой переменной, вывод нескольких точек данных для иллюстрации структуры данных.
- Провести описательный анализ данных. Сделать выводы. Этот этап включает в себя определение шкалы измерения каждого признака, выявление аномальных значений, визуализацию распределения каждого признака, при необходимости - проверка на нормальность, построение коррелограмм и совместных распределений каждого признака

с целевой переменной, выявление коррелированных признаков и признаков, не несущих информации для данной задачи.

- Сделать выводы

Методические указания:

- Пояснительная записка выполняется в виде текстового документа и должна включать в себя: титульный лист, текстовое описание проблемы, ссылку на публично доступный репозиторий с полным кодом выполнения работы, по необходимости пример кода для каждого этапа работы, текстовые выводы по каждому этапу и сформулированное заключение с результатами работы и их интерпретация.
- Все пояснения, выводы и замечания, на которые необходимо обратить внимание должны присутствовать в работе в виде ячеек документации либо (менее предпочтительно) программных комментариев.
- Работа должна выполняться студентом самостоятельно и индивидуально.
- Оценка качества моделирования должна производиться с использованием определенных метрик. Их выбор должен быть описан и обоснован до начала моделирования. Плюсом работы является широкий набор метрик эффективности моделей.
- Отчет работы производится в формате презентации. Слушатели (включая преподавателя) могут задавать вопросы представляющему свою работу студенту. Регламент презентации - 5 минут на выступление, 2 минуты на вопросы.

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

Комплект лицензионного программного обеспечения:

1. Python 3.8
2. Jupyter Notebook (Windows 10)
3. Kaspersky

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/>

Сертифицированные программные и аппаратные средства защиты информации:

1. не предусмотрены

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

1. **Компьютерный класс** для проведения учебных занятий, предусмотренных программой, в том числе групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенный оборудованием и техническими средствами обучения: мебель аудиторная (столы, стулья, доска аудиторная), персональные компьютеры, набор демонстрационного оборудования (проектор, экран)
2. **Учебная аудитория** для проведения учебных занятий, предусмотренных программой, в том числе групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная оборудованием и техническими средствами обучения: мебель аудиторная (столы, стулья, доска аудиторная), набор демонстрационного оборудования (проектор, экран)