Приложение к ОПОП
«Математические методы в информационных технологиях»

СПбГЭТУ «ЛЭТИ»
ПЕРВЫЙ ЭЛЕКТРОТЕХНИЧЕСКИЙ

МИНОБРНАУКИ РОССИИ

федеральное государственное автономное образовательное учреждение высшего образования
**«Санкт-Петербургский государственный электротехнический университет
«ЛЭТИ» им. В.И.Ульянова (Ленина)»
(СПбГЭТУ «ЛЭТИ»)**

РАБОЧАЯ ПРОГРАММА

дисциплины

«БОЛЬШИЕ ДАННЫЕ»

для подготовки бакалавров

по направлению

01.03.02 «Прикладная математика и информатика»

по профилю

«Математические методы в информационных технологиях»

Санкт-Петербург

2023

ЛИСТ СОГЛАСОВАНИЯ

Разработчики:
заведующий кафедрой, д.пед.н., доцент Поздняков С.Н.

Рабочая программа рассмотрена и одобрена на заседании кафедры АМ 12.01.2023, протокол № 6

Рабочая программа рассмотрена и одобрена учебно-методической комиссией ФКТИ, 16.02.2023, протокол № 2

Согласовано в ИС ИОТ
Начальник ОМОЛА Загороднюк О.В.

1 СТРУКТУРА ДИСЦИПЛИНЫ

Обеспечивающий факультет	ФКТИ
Обеспечивающая кафедра	ИС
Общая трудоемкость (ЗЕТ)	3
Курс	4
Семестр	7
Виды занятий	
Лекции (академ. часов)	34
Практические занятия (академ. часов)	34
Иная контактная работа (академ. часов)	1
Все контактные часы (академ. часов)	69
Самостоятельная работа, включая часы на контроль (академ. часов)	39
Всего (академ. часов)	108
Вид промежуточной аттестации	
Экзамен (курс)	4

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«БОЛЬШИЕ ДАННЫЕ»

Данный курс обеспечивает теоретические и практические знания в области методов и инструментов работы с Большими данными. Программа курса включает в себя изучение понятия Больших данных, особенностей работы с ними и средств обеспечивающих их масштабируемый анализ. В рамках дисциплины рассматриваются средства для работы с данными различного вида: структурированными, пвседоструктурированными, не-структурированными, потоковыми, распределенными и другими. Изучаются основные парадигмы распределенной обработки данных, такие как MapReduce, лямбда-архитектуры и другие, а также особенности методов анализа применяемых к Большим данным.

Рассматриваются и сравниваются два основных подходов: централизованный анализ, предполагающий предварительный сбор данных в единое хранилище, и федеративный анализ, предполагающий выполнение анализа непосредственно на источниках данных, с последующей агрегацией результатов. В рамках централизованного анализа рассматриваются три поколения платформ анализа данных: хранилища данных, ”озера” данных и потоковая обработка данных.

Приобретаются практические навыки анализа Больших данных. Все занятия дисциплины подкреплены примерами.

SUBJECT SUMMARY

«BIG DATA»

This course provides theoretical and practical knowledge in the field of methods and tools for working with Big Data. The course program includes the study of the concept of Big Data, the features of working with them and the tools that provide their scalable analysis. Within the framework of the discipline, tools for working

with data of various types are considered: structured, pseudo-structured, non-structured, streaming, distributed, and others. We study the main paradigms of distributed data processing, such as MapReduce, lambda architectures and others, as well as the features of analysis methods applied to Big Data.

Two main approaches are considered and compared: centralized analysis, which involves the preliminary collection of data in a single repository, and federated analysis, which involves performing the analysis directly on data sources, followed by aggregation of the results. Centralized analysis focuses on three generations of data analysis platforms: data warehouses, data lakes, and streaming data processing.

Acquire practical skills in Big Data analysis. All lessons of the discipline are supported by examples.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Цель дисциплины -формирование знаний, умений и навыков владения средствами и методами обработки и анализа Больших данных.

2. Задачами дисциплины являются:
-приобретение знаний по базовым методам и алгоритмам обработки и анализа Больших данных и их усовершенствования для выполнения в параллельной и распределенной среде;
-формирование умений и практических навыков разработки алгоритмического и программного обеспечения методов анализа Больших данных;
-освоение навыков применения методов и алгоритмов анализа больших данных.

3. Получение знаний по базовым методам построения платформ Больших данных и алгоритмам обработки и анализа Больших данных и их усовершенствования для выполнения в параллельной и распределенной среде.

4. Формирование умений и практических навыков разработки алгоритмического и программного обеспечения методов анализа Больших данных.

5. Освоение навыков применения методов и алгоритмов анализа Больших данных.

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе ранее освоенных дисциплин учебного плана:

1. «Математические основы машинного обучения»

2. «Производственная практика (технологическая (проектно-технологическая) практика)»

и обеспечивает изучение последующих дисциплин:

1. «Производственная практика (научно-исследовательская работа)»

2. «Производственная практика (преддипломная практика)»

3.3 Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы

В результате освоения образовательной программы обучающийся должен достичь следующие результаты обучения по дисциплине:

Код компетенции/ индикатора компетенции	**Наименование компетенции/индикатора компетенции**
ПК-0	Способен разрабатывать информационные модели и применять их для решения задач профессиональной деятельности
ПК-0.2	*Создает и модифицирует информационные модели для решения задач профессиональной деятельности*

4 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

4.1 Содержание разделов дисциплины

4.1.1 Наименование тем и часы на все виды нагрузки

№ п/п	Наименование темы дисциплины	Лек, ач	Пр, ач	ИКР, ач	СР, ач
1	Введение в курс	1			
2	Поколения платформ данных	4			
3	Распределенная обработка данных	6	9		7
4	Федеративное обучение	8	10		11
5	Хранение Больших данных	4	5		0
6	Обработка потоковых данных	2			7
7	Алгоритмы анализа Больших данных	8	10		14
8	Заключение	1		1	
	Итого, ач	34	34	1	39
	Из них ач на контроль	0	0	0	35
	Общая трудоемкость освоения, ач/зе	108/3			

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Введение в курс	Общая информация о структуре курса, изучаемых технологиях и методах.
2	Поколения платформ данных	Три поколения платформ Больщих данных: Хранилища данных, Озера данных, Потоковые системы. Основные принципы их построения и отличия друг от друга.
3	Распределенная обработка данных	Основные понятия распределенных систем. CAP теорема. Требования к распределенным системам со стороны задач обработки данных. Распределенное обучение на GPU. Концепция MapReduce
4	Федеративное обучение	Технология федеративного обучения. Назначения и возможности. Классификация системе федеративного обучения. Типы распределенных данных. Основные фреймворки федеративного обучения. Основные алгоритмы.
5	Хранение Больших данных	Структурированные, полуструктурированные и неструктурированные данные. ACID требования. Способы горизонтального масштабирования. SQL, noSQL и newSQL системы хранения.

№ п/п	Наименование темы дисциплины	Содержание
6	Обработка потоковых данных	Потоковые данные. Системы обработки потоковых сообщений. Лямбда архитектура для обработки потоковых данных. Гамма архитектура для обработки потоковых данных.
7	Алгоритмы анализа Больших данных	Особенности алгоритмов обработки Больших данных. Принципы построения алгоритмов обработки Больших данных. Основные алгоритмы обработки Больших данных.
8	Заключение	Подведение итогов. Обобщение пройденного материала

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Анализ и постановка задачи машинного обучения на выбранном наборе данных	10
2. Последовательное выполнение алгоритма машинного обучения	12
3. Масштабированное выполнение алгоритма машинного обучения	12
Итого	34

4.4 Курсовое проектирование

Курсовая работа (проект) не предусмотрены.

4.5 Реферат

Реферат не предусмотрен.

4.6 Индивидуальное домашнее задание

Индивидуальное домашнее задание не предусмотрено.

4.7 Доклад

Студент должен подготовить доклад в виде выступления с презентацией.

Оценка выставляется исходя из:

- своевременности присланного доклада;

- качества презентации (соответствие структуре, логической связанности, полноты, качеству слайдов);

- качества выступления (выразительности, связанности и четкости рассказа);

- ответов на вопросы (3 вопроса).

Время доклада 10 мин (это 7-10 слайдов). До выступления доклад должен быть одобрен преподавателем, поэтому презентации должны присылаться заранее с расчетом, что нужно будет исправлять замечания.

Последний срок посылки первой версии доклада воскресенье до конца дня.

Темы и структуры докладов:

Тема I. Системы распределенной обработки данных

Структура доклада:

1.Общая информация. Разработчик. Лицензия. Год выпуска. Текущая версия. Дата версии.

2.Назначение системы. Кейсы применения.

3.Архитектура системы. Модель распределенных вычислений.

4.Реализация масштабирования, производительности (”проблема отстающего”, (slow node problem)).

5. Реализация прозрачности.

6. Реализация универсальности.

7. Реализация отказоустойчивости.

8.Интерфейс взаимодействия (API, протокол взаимодействия).

9.Достоинства и недостатки системы.

10.Примеры применения.

11.Вывод.

Тема II. Фреймворки федеративного обучения

Структура доклада:

1.Общая информация: разработчик, ссылка, версия, язык, лицензия.

2.Архитектура. Протоколы коммуникаций. Используемые фреймворки.

3.API для работы с FL.

4.Топология: централизованная/децентрализованная, cross-silo, cross-devices.

5.Разделение данных: Horizontal, Vertical, Hybrid.

6.Реализованные ML алгоритмы.

7.Механизмы защиты.

8. Выводы.

Тема III. Системы хранения Больших данных

Структура доклада:

1. Общая информация. Разработчик. Лицензия. Год выпуска. Текущая версия. Дата версии.

2. Назначение системы. Кейсы применения.

3. Модель хранения данных.

4. Архитектура системы.

5. Реализация: масштабирования, доступности, отказоустойчивость, согласованность, и др.

6. Информационные принципы взаимодействия (API, протокол взаимодействия).

7. Основные характеристики (объем данных, форматы данных, время обработки, и т.п.).

8. Достоинства и недостатки системы.

9. Примеры применения.

10. Вывод.

4.8 Кейс

Кейс не предусмотрен.

4.9 Организация и учебно-методическое обеспечение самостоятельной работы

Изучение дисциплины сопровождается самостоятельной работой студентов с рекомендованными преподавателем литературными источниками и информационными ресурсами сети Интернет.

Планирование времени для изучения дисциплины осуществляется на весь период обучения, предусматривая при этом регулярное повторение пройденного материала. Обучающимся, в рамках внеаудиторной самостоятельной работы, необходимо регулярно дополнять сведениями из литературных источников материал, законспектированный на лекциях. При этом на основе изучения рекомендованной литературы целесообразно составить конспект основных положений, терминов и определений, необходимых для освоения разделов учебной дисциплины.

Особое место уделяется консультированию, как одной из форм обучения и контроля самостоятельной работы. Консультирование предполагает особым образом организованное взаимодействие между преподавателем и студентами, при этом предполагается, что консультант либо знает готовое решение, которое он может предписать консультируемому, либо он владеет способами деятель-

ности, которые указывают путь решения проблемы.

Самостоятельное изучение студентами теоретических основ дисциплины обеспечено необходимыми учебно-методическими материалами (учебники, учебные пособия, конспект лекций и т.п.), выполненными в печатном или электронном виде.

Изучение студентами дисциплины сопровождается проведением регулярных консультаций преподавателей, обеспечивающих практические занятия по дисциплине, за счет бюджета времени, отводимого на консультации (внеаудиторные занятия, относящиеся к разделу «Самостоятельные часы для изучения дисциплины»).

Текущая СРС	**Примерная трудоемкость, ач**
Работа с лекционным материалом, с учебной литературой	0
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	0
Выполнение домашних заданий, домашних контрольных работ	0
Подготовка к лабораторным работам, к практическим и семинарским занятиям	0
Подготовка к контрольным работам, коллоквиумам	0
Выполнение расчетно-графических работ	0
Выполнение курсового проекта или курсовой работы	0
Поиск, изучение и презентация информации по заданной проблеме, анализ научных публикаций по заданной теме	4
Работа над междисциплинарным проектом	0
Анализ данных по заданной теме, выполнение расчетов, составление схем и моделей, на основе собранных данных	0
Подготовка к зачету, дифференцированному зачету, экзамену	35
ИТОГО СРС	39

5 Учебно-методическое обеспечение дисциплины

5.1 Перечень основной и дополнительной литературы, необходимой для освоения дисциплины

№ п/п	Название, библиографическое описание	К-во экз. в библ.
Основная литература		
1	Интеллектуальный анализ распределенных данных на базе облачных вычислений [Текст] / [М.С. Куприянов [и др.], 2011. -147 с.	9
2	Интеллектуальный анализ данных в распределенных системах [Текст] : [монография] / [М. С. Куприянов, И. И. Холод, З. А. Каршиев, И. А. Голубев], 2012. -108, [1] с.	9
Дополнительная литература		
1	Методы и модели анализа данных : OLAP и Data Mining [Текст] : учеб. пособие по специальности 071900 информ. системы и технологии" направления 654700 "Информ. системы" / А. А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод, 2004. -336 с.	67
2	Технологии анализа данных [Текст] : Data Mining, Visual Mining, Text Minning, OLAP : учеб. пособие по специальности 071900 "информац. системы и технологии" направления 654700 "Информационные системы" / А.А. Барсегян [и др.], 2007. -VIII, 375 с.	43
3	Методы оперативного и интеллектуального анализа данных [Текст] : метод. указания к лаб. работам / Санкт-Петербургский государственный электротехнический университет им. В.И. Ульянова (Ленина) "ЛЭТИ", 2008. -32 с.	85

5.2 Перечень ресурсов информационно-телекоммуникационной сети «Интернет», используемых при освоении дисциплины

№ п/п	Электронный адрес
1	Что такое большие данные?https://www.oracle.com/cis/big-data/what-is-big-data/#:~:text=%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5%20%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5%20%E2%80%94%20%D1%8D%D1%82%D0%BE%20%D1%80%D0%B0%D0%B7%D0%BD%D0%BE%D0%BE%D0%B1%D1%80%D0%B0%D0%B7%D0%BD%D1%8B%D0%B5%20%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5,%D1%81%D0%BA%D0%BE%D1%80%D0%BE%D1%81%D1%82%D1%8C%20%D0%BF%D0%BE%D1%81%D1%82%D1%83%D0%BF%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F%20%D0%B8%20%D0%B1%D0%BE%D0%BB%D1%8C%D1%88%D0%BE%D0%B9%20%D0%BE%D0%B1%D1%8A%D0%B5%D0%BC.

№ п/п	Электронный адрес
2	Большие данные (Big Data)https://www.tadviser.ru/index.php/%D0%A1%D1%82%D0%B0%D1%82%D1%8C%D1%8F:%D0%91%D0%BE%D0%BB%D1%8C%D1%88%D0%B8%D0%B5_%D0%B4%D0%B0%D0%BD%D0%BD%D1%8B%D0%B5_(Big_Data)
3	Что такое большие данные?https://www.sap.com/cis/insights/what-is-big-data.html

5.3 Адрес сайта курса

Адрес сайта курса: https://vec.etu.ru/moodle/course/view.php?id=11863

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Большие данные» формой промежуточной аттестации является экзамен. Оценивание качества освоения дисциплины производится с использованием рейтинговой системы.

Экзамен

Оценка	Количество баллов	Описание
Неудовлетворительно	0 – 30	теоретическое содержание курса не освоено, необходимые практически навыки и умения не сформированы, выполненные учебные задания содержат грубые ошибки, дополнительная самостоятельная работа над курсом не приведет к существенному повышению качества выполнения учебных заданий
Удовлетворительно	31 – 60	теоретическое содержание курса освоено частично, но пробелы не носят существенного характера, необходимые практические навыки и умения работы с освоенным материалом в основном сформированы, большинство предусмотренных программой обучения учебных заданий выполнено, некоторые из выполненных заданий содержат ошибки
Хорошо	61 – 80	теоретическое содержание курса освоено полностью, без пробелов, некоторые практические навыки и умения сформированы недостаточно, все предусмотренные программой обучения учебные задания выполнены, качество выполнения ни одного из них не оценено минимальным числом баллов, некоторые виды заданий выполнены с ошибками
Отлично	81 – 100	теоретическое содержание курса освоено полностью, без пробелов, необходимые практические навыки и умения сформированы, все предусмотренные программой обучения учебные задания выполнены, качество их выполнения оценено количеством баллов, близким к максимальному

Особенности допуска

Подготовка и выступление с тремя докладами/презентациями. Допуск к экзамену при оценке за практическую часть ≥ 30.

6.2 Оценочные материалы для проведения текущего контроля и промежуточной аттестации обучающихся по дисциплине

Вопросы к экзамену

№ п/п	Описание
1	Информационно-аналитические системы. Назначение. Структура
2	Противоречия между OLTP и ИАС
3	Хранилища данных. Назначение. Архитектура
4	Перенос данных. ETL процесс
5	Очистка данных
6	Типы хранилищ данных
7	Виды анализа данных
8	Многомерный анализ данных
9	Концепция OLAP
10	Неструктурированные и псевдоструктурированные данные
11	Озера данных
12	NoSQL БД
13	NewSQL БД
14	Потоковые данные
15	Лямбда архитектура
16	Концепция MapReduce
17	Федеративное обучение
18	Алгоритм FedAvg
19	Алгоритм SecAgg
20	Особенности алгоритмов Больших данных

Форма билета

Министерство науки и высшего образования Российской Федерации

ФГАОУ ВО «Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» имени В.И. Ульянова (Ленина)»

_____

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

Дисциплина **Большие данные**

1. Информационно-аналитические системы. Назначение. Структура

2. Федеративное обучение. Преимущества и недостатки

УТВЕРЖДАЮ

Заведующий кафедрой В.В. Цехановский

Весь комплект контрольно-измерительных материалов для проверки сформированности компетенции (индикатора компетенции) размещен в закрытой части по адресу, указанному в п. 5.3

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
1	Распределенная обработка данных	
2		
3		
4		Доклад / Презентация
5	Федеративное обучение	
6		
7		
8		Доклад / Презентация
9	Хранение Больших данных	
10		
11		
12		
13		Доклад / Презентация

6.4 Методика текущего контроля

Оценка по дисциплине формируется из:

- оценки за теоретическую часть (максимум 40 баллов);

- оценки за практическую часть (минимум 30, максимум 60 баллов).

Оценка за теоретическую часть может быть получена:

- за ответы на вопросы на лекциях;

- за доклад (максимум 10 баллов);

- за оппонирование докладов (максимум 5 баллов);

- за экзамен (максимум 10 баллов за вопрос).

Оценка за практическую часть может быть получена за доклады и выполнение заданий:

- доклад 1 – максимум 20 баллов

- доклад 2 – максимум 40 баллов:

последовательный алгоритм (10 баллов);

масштабированный алгоритм (10 баллов);

Оценка за доклад выставляется исходя из:

- своевременности присланного доклада;

- качества презентации (соответствие структуре, логической связанности, полноты, качеству слайдов);

- качества выступления (выразительности, связанности и четкости рассказа);

- ответов на вопросы (3 вопроса).

Время доклада 10 мин (это 7-10 слайдов). До выступления доклад должен быть одобрен преподавателем, поэтому презентации должны присылаться заранее с расчетом, что нужно будет исправлять замечания.

20 дополнительных баллов за выполнения практической части до 01.05.

самостоятельной работы студентов

Контроль самостоятельной работы студентов осуществляется на лекционных и практических занятиях студентов по методикам, описанным выше.

7 Описание информационных технологий и материально-технической базы

Тип занятий	Тип помещения	Требования к помещению	Требования к программному обеспечению
Лекция	Лекционная аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, проектор, экран, ПК или ноутбук	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Практические занятия	Помещение для выполнения практических заданий	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя. Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к DataSpere в Яндекс.Облако	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Самостоятельная работа	Помещение для самостоятельной работа	Оснащено компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду университета.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше

8 Адаптация рабочей программы для лиц с ОВЗ

Адаптированная программа разрабатывается при наличии заявления со стороны обучающегося (родителей, законных представителей) и медицинских показаний (рекомендациями психолого-медико-педагогической комиссии). Для инвалидов адаптированная образовательная программа разрабатывается в соответствии с индивидуальной программой реабилитации.

ЛИСТ РЕГИСТРАЦИИ ИЗМЕНЕНИЙ

№ п/п	Дата	Изменение	Дата и номер протокола заседания УМК	Автор	Начальник ОМОЛА