МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное бюджетное образовательное учреждение высшего образования

«ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

(ФГБОУ ВО «ИГУ»)

Институт математики и информационных технологий

«УТВЕРЖДАЮ»
Директор ИМИТ ИГУ
_____ М. В. Фалалеев
«10» мая 2023 г.

Рабочая программа дисциплины (модуля)

Б1.В.07.06 Анализ распределенных данных

Направление подготовки: 01.04.02 Прикладная математика и информатика

Направленность (профиль) подготовки: Семантические технологии и многоагентные системы

Квалификация выпускника: магистр

Форма обучения: очная

Иркутск 2023 г.

2 АННОТАЦИЯ ДИСЦИПЛИНЫ

«АНАЛИЗ РАСПРЕДЕЛЕННЫХ ДАННЫХ»

Дисциплина посвящена подходам и методам анализа данных формируемых и хранящихся на распределенных узлах связанных между собой вычислительной сетью. 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.

SUBJECT SUMMARY

«ANALYSIS OF DISTRIBUTED DATA»

The discipline is devoted to the approaches and methods of analyzing data generated and stored on distributed nodes interconnected by a computer network. Two main approaches are considered and compared: centralized analysis, which presupposes the preliminary collection of data into a single repository, and federated analysis, which involves performing analysis directly on data sources, with subsequent aggregation of the results. Centralized analysis looks at three generations of data analysis platforms: data warehouses, data lakes, and data streaming.

3 ОБЩИЕ ПОЛОЖЕНИЯ

3.1 Цели и задачи дисциплины

1. Изучение подходов и методов анализа данных, формируемых и хранящихся на распределенных узлах, связанных между собой вычислительной сетью, и приобретение навыков использования полученных знаний в профессиональной деятельности.

2. Ознакомиться с подходами и инструментами анализа распределенных данных

Ознакомиться с методами и инструментами федеративного обучения

Получить навыки использования инструментов федеративного обучения

3. Знание подходов и методов анализа распределенных данных

4. Умения выбирать методы и инструменты для анализа распределенных данных

5. Навыки использования инструментов федеративного обучения

3.2 Место дисциплины в структуре ОПОП

Дисциплина изучается на основе знаний, полученных при освоении программы бакалавриата или специалитета.

и обеспечивает изучение последующих дисциплин:

1. «Аналитические информационные системы»

2. «Семантический Web»

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-12	Способен разрабатывать и применять методы и алгоритмы машинного обучения для решения задач
ПК-12.1	*Ставит задачи по разработке или совершенствованию методов и алгоритмов для решения комплекса задач предметной области*
ПК-12.2	*Разрабатывает унифицированные и обновляемые методологии описания, сбора и разметки данных, а также механизмы контроля за соблюдением указанных методологий*

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	КО, ач	СР, ач
1	Введение в дисциплину	1			2
2	Поколения платформ анализа данных. Хранилища данных.	1			2
3	Неструктурированные данные. Озера данных	2			2
4	Потоковые данные	2		1	1
5	Федеративное обучение	4	8		39
	Итого, ач	10	8	8	46
	Из них ач на контроль	0	0	0	0
	Общая трудоемкость освоения, ач/зе	72/2			

4.1.2 Содержание

№ п/п	Наименование темы дисциплины	Содержание
1	Введение в дисциплину	Структура, цели, задачи и содержание дисциплины.

2	Поколения платформ анализа данных. Хранилища данных.	Три поколения платформ анализа распределенных данных. Концепция хранилища данных. ETL процесс. Проблемы интеграции данных из распределенных источников.
3	Неструктурированные данные. Озера данных	Неструктурированные и полу структурированные данные. Концепция ”озер” данных. Средства распределенного анализа данных. Концепция MapReduce. Основные инструменты распределенного анализа данных.
4	Потоковые данные	Понятие потоковых данных. Проблемы обработки потоковых данных. Свойства систем анализа потоковых данных. Лямбда-архитектура. Капа-архитектура.
5	Федеративное обучение	Проблемы централизованного анализа данных. Федеративное обучение. Виды систем федеративного обучения. Проблемы построения систем федеративного обучения. Основные инструменты федеративного обучения. Основные алгоритмы федеративного обучения.

4.2 Перечень лабораторных работ

Лабораторные работы не предусмотрены.

4.3 Перечень практических занятий

Наименование практических занятий	Количество ауд. часов
1. Изучение инструментов федеративного обучения	8
Итого	8

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

Темы для подготовки докладов:

1. TensorFlow Federated (TFF) from Google Inc (USA)

2. Federated AI Technology Enabler (FATE) from Webank's AI department (China)

3. Paddle Federated Learning (PFL) from Baidu (China)

4. FedML

5. PySyft from open community OpenMined

6. FEDn 0.2.3 from Scaleout Systems (Sweden)

7. *Nvidia Clara Train SDK (3.1) from Nvidia (USA)*

8. *IBM FL (1.0.3) from IBM (USA)*

9. Flower framework

10. Fedlearner https://github.com/bytedance/fedlearner.

11. Sherpa.ai Federated Learning and Differential Privacy Framework

12. HP Swarm learning

План доклада по выполнению анализа данных:

1. Подходы к анализу выбранных данных ;

2. Результаты анализа данных.

Результаты выполнения работы: Презентация и отчет.

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 ин-

формационными

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 особымобразом 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, онлайн-версия

курса), выполненными в печатном или электронном виде.

Текущая СРС	**Примерная трудоемкость, ач**
Работа с лекционным материалом, с учебной литературой	5
Опережающая самостоятельная работа (изучение нового материала до его изложения на занятиях)	0
Самостоятельное изучение разделов дисциплины	5
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	36
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	8
ИТОГО СРС	54

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
Дополнительная литература		
1	Цехановский, Владислав Владимирович . Интеллектуальный анализ данных [Текст] : учеб. пособие / В. В. Цехановский, В. Д. Чертовской, 2019. -55 с.	23
2	Мостеллер, Фредерик. Анализ данных и регрессия [Текст] : [в 2 вып.]. Вып. 2 / пер. с англ. Б. Л. Розовского ; под ред. и с предисл. Ю. П. Адлера, 1982. -235, [2] с.	23
3	Мостеллер, Фредерик. Анализ данных и регрессия [Текст] : [в 2 вып.]. Вып. 1 / пер. с англ. Ю. Н. Благовещенского ; под ред. и с предисл. Ю. П. Адлера, 1982. -318, [1] с.	24

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

№ п/п	Электронный адрес
1	DATA SCIENCE. Федеративное обучение. https://datascience.eu/ru/%D0%BC%D0%B0%D1%8%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5/%D1%8%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5/

6 Критерии оценивания и оценочные материалы

6.1 Критерии оценивания

Для дисциплины «Анализ распределенных данных» формой промежуточной аттестации является дифф. зачет. Оценивание качества освоения дисциплины производится с использованием рейтинговой системы.

Дифференцированный зачет

Оценка	Количество баллов	Описание
Неудовлетворительно	0 – 51	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	52 – 67	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	68 – 84	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	85 – 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

Допуск к дифференцированному зачету -сумма баллов не менее 51 баллов по всем видам работ.

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

Примерные вопросы к дифф.зачету

№ п/п	Описание
1	Поколения платформ анализа данных. Хранилища данных
2	Неструктурированные и полу структурированные данные.
3	Концепция MapReduce. Основные инструменты распределенного анализа данных.
4	Понятие потоковых данных. Проблемы обработки потоковых данных.
5	Проблемы централизованного анализа данных. Федеративное обучение
6	Основные инструменты федеративного обучения.
7	Основные алгоритмы федеративного обучения.

6.3 График текущего контроля успеваемости

Неделя	Темы занятий	Вид контроля
10	Федеративное обучение Поколения платформ анализа данных. Хранилища данных. Потоковые данные Неструктурированные данные. Озера данных	
11		
12		
13		
14		
15		
16		Доклад / Презентация

6.4 Методика текущего контроля

Оценка по дисциплине формируется из:

–оценку за теоретическую часть (максимум 40 баллов);

–оценку за практическую часть (максимум 60 баллов).

Оценка за теоретическую часть может быть получена:

-за ответы на вопросы на лекциях (максимум 20 баллов);

-за доклад (максимум 10 баллов за 1 доклад)

-за тест (максимум 20 баллов).

Оценка за практическую часть выставляется за выполнение заданий практическим занятиям и доклады по ним (максимум 30 баллов за задание).

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, компьютер или ноутбук, проектор, экран, маркерная доска.	1) Windows XP и выше; 2) Microsoft Office 2007 и выше
Практические занятия	Аудитория	Количество посадочных мест – в соответствии с контингентом, рабочее место преподавателя, компьютер или ноутбук, проектор, экран, маркерная доска.	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.