АННОТАЦИЯ ДИСЦИПЛИНЫ

«Библиотеки и фреймворки для кластерных технологий обработки больших данных»

Дисциплина «Библиотеки и фреймворки для кластерных технологий обработки больших данных» является частью программы магистратуры « Технологии искусственного интеллекта в социальных и экономических системах» по направлению «09.04.01 Информатика и вычислительная техника».

Цели и задачи дисциплины

Формирование комплекса знаний, умений и навыков в области построения распределенных кластерных систем для обработки больших данных средствами современных языков программирования..

Изучаемые объекты дисциплины

Большие данные; распределенные кластерные системы; параллельные вычисления; язык программирования Python; модули и библиотеки научных вычислений; модули и библиотеки обработки данных..

Объем и виды учебной работы

Вид учебной работы	Всего часов	Распределение по семестрам в часах
		Номер семестра
		3
1. Проведение учебных занятий (включая проведе-ние текущего контроля успеваемости) в форме: 1.1. Контактная аудиторная работа, из них:	54	54
- лекции (Л)	18	18
- лабораторные работы (ЛР)	18	18
- практические занятия, семинары и (или) другие виды занятий семинарского типа (ПЗ)	16	16
- контроль самостоятельной работы (КСР)	2	2
- контрольная работа		
1.2. Самостоятельная работа студентов (СРС)	90	90
2. Промежуточная аттестация		
Экзамен		
Дифференцированный зачет	9	9
Зачет		
Курсовой проект (КП)		
Курсовая работа (КР)		
Общая трудоемкость дисциплины	144	144

Краткое содержание дисциплины

Наименование разделов дисциплины с кратким содержанием	Объем аудиторных занятий по видам в часах			Объем внеаудиторных занятий по видам в часах
	Л	ЛР	ПЗ	СРС
3-й семестр				
Инструменты распределенного решения научных задач	2	4	4	20
Наукоемкие вычислительные задачи. Понятие «добровольные вычисления» в контексте грид-вычислений. BOINC. World Community Grid. SETI@Home. Folding@Home. Rosetta@home. Преимущества и критика проектов.				
Сбор и хранение больших данных	4	4	4	20
Понятие больших данных. Понятие «озера данных» (Data Lake). Мультимодальные большие данные. Распределенные хранилища. SQL- и NoSQL-подходы. Колоночные базы данных. Графовые базы данных. Распределенные файловые системы. HDFS. Cassandra. Greenplum. ClickHouse.				
Использование кластерных технологий в машинном обучении	4	4	4	20
Задачи машинного обучения, требующие больших данных. Обучение ИНС на кластерах. CUDA-кластеры. PyTorch. Onyx. TensorFlow. Scikit-Learn. Большие языковые модели (LLMs) . Трансформеры (BERT, GPT). Мультимодальные генеративные сети (DALL-E, Stable Diffusion, Whisper).				
Введение в кластерные технологии	4	2	0	10
Понятие параллельных вычислений. Понятие вычислительного кластера. ЦОД. Подходы к архитектуре распределенных приложений. DAG-топология. Потоковая обработка данных. Масштабируемость. Отказоустройчивость.				
Инструменты распределенного анализа больших данных	4	4	4	20
Задачи распределенного анализа. ETL-процедуры в кластере. Интеллектуальный анализ данных (data mining). Apache Spark. Hadoop. Zookeeper. Samza. Flink. Kafka. Elasticsearch.				
ИТОГО по 3-му семестру	18	18	16	90
ИТОГО по дисциплине	18	18	16	90