Аннотация рабочей программы дисциплины

Б1.В.12 «Технологии обработки больших данных»

Курс 4 Семестр 7 Количество з.е. 4
Объем трудоемкости: 4 зачетных единиц (144 ч., из них – 68 час. аудиторной нагрузки: лекционных 34ч., лабораторных работ - 34 ч., 36 часов самостоятельной работы, 4 часа КСР, 0,3 часа ИКР.), форма контроля – экзамен.

Цель освоения дисциплины Цель дисциплины - Изучение принципов обработки больших данных, технологий распределенных вычислений, облачных платформ и инструментов анализа данных.

Задачи дисциплины:
- Изучение архитектурных решений для работы с Big Data.
- Освоение методов обработки структурированных и неструктурированных данных.
- Применение распределенных вычислений (Hadoop, Spark).
- Разработка алгоритмов анализа данных в распределенных средах.
- Использование облачных платформ для обработки больших данных.

Требования к знаниям и навыкам:
- Умение работать с распределенными системами (Hadoop, Spark).
- Опыт обработки данных в облачных средах (Yandex Cloud).
- Навыки анализа данных с помощью Python (Pandas, PySpark, Dask).
- Понимание архитектуры Big Data-решений.

Место дисциплины в структуре ООП ВО:

Дисциплина «Технологии обработки больших данных» относится к Блок 1 Дисциплины, часть, формируемая участниками образовательного процесса.

Дисциплина изучается в 7-м семестре. Для успешного освоения необходимы знания, полученные в дисциплинах: «Алгебра и введение в тензорный анализ», «Теория вероятностей и математическая статистка», «Подготовка данных машинного обучения», «Технологии управления данными NoSQL», «Многомерный статистический анализ», и «Машинное обучение», «Программирование».

Преподавание ведется в виде лекций и лабораторных занятий с использованием интерактивных методов. Лабораторные работы направлены на практическое освоение методов и инструментов классификации на реальных данных.

Дисциплина формирует компетенции, необходимые для выполнения выпускной квалификационной работы и профессиональной деятельности в области вычислительных технологий.

Результаты обучения (знания, умения, опыт, компетенции):

BD-3	**Способен организовывать хранения данных, выбирая адекватные технологические решения**

BD-3.1	Разрабатывает, отлаживает и тестирует прикладные решения с элементами ИИ с применением различных технологий хранения структурированных данных, оценивает качество. Пишет аналитические запросы к данным и анализирует план запроса. Умеет создавать представления, хранимые процедуры, функции и триггеры. Знание типов СУБД: реляционные, NoSQL, колоночные, документные - Архитектура распределенных систем хранения - Принципы ACID, CAP-теорема - Методы индексации и партиционирования - Принципы транзакционности Уметь: Выбирать тип СУБД под задачи ИИ - Проектировать схемы данных для ML-моделей - Анализировать и оптимизировать планы запросов - Разрабатывать сложные аналитические запросы - Создавать database objects (views, procedures, functions) Владение SQL (оконные функции, CTE, сложные джойны) - Навык чтения explain plan - Оптимизация запросов через индексы - Создание хранимых процедур для ETL - Работа с триггерами для поддержания целостности
BD-4	**Способен применять различные модели и (или) технологии обработки данных**
BD-4.1	Осуществляет выбор технологий обработки больших данных, приемлемых для создания прикладной системы ИИ с заданными требованиями Способен организовывать распределенное хранилище и параллельную обработку на базе современных технологий (Hadoop, Spark) больших данных: Знать: Архитектуру Hadoop ecosystem (HDFS, YARN) - Модели вычислений: MapReduce, DAG - Принципы RDD и DataFrame в Spark – принципы Стриминговой обработки vs batch processing - Паттерны Lambda/Kappa архитектур Уметь: Выбирать стек технологий под задачи ИИ - Проектировать распределенные ETL-пайплайны - Оптимизировать производительность Spark-приложений - Организовывать шардирование и репликацию данных Владеть PySpark API - Настройка и администрирование Hadoop/Spark кластеров - Оптимизация через партиционирование, кэширование - Мониторинг производительности распределенных систем
BD-5	**Способен применять технологии организации инфраструктуры БД**
BD-5.1	Осуществляет выбор направления вспомогательных технологических решений для формирования единого стека работы с большими данными для решения поставленной задачи. Руководит проектами по организации инфраструктуры БД: Знать: принципы построения data lake, data warehouse - Методы оркестрации пайплайнов (Airflow, Prefect) - CI/CD для data projects - Мониторинг и observability в data-системах - Методологии управления data projects. Уметь: Формировать единый технологический стек - Управлять жизненным циклом data infrastructure - Выбирать инструменты

	мониторинга и оркестрации - Оценивать TCO инфраструктурных решений. Владеть навыками проектного управления в data-проектах - Составление ТЗ на инфраструктуру - Ведение технической документации - Оценка рисков инфраструктурных решений.
ML-1	**Способен применять знания об истории развития и трендах современного ИИ для формулирования корректных постановок задач и поиска перспективных способов решения проблем с помощью ИИ**
ML-1.2	Определяет тенденции развития, оценивает новизну и практическую значимость своих решений с точки зрения современного искусственного интеллекта. Проектирует и внедряет комплексные пайплайны предварительной обработки данных с использованием современных методов ИИ, автоматизации и feature engineering в различных предметных областях. Знать: современные архитектуры ML-моделей (трансформеры, GAN, RL) - Методы feature engineering и feature selection - MLOps принципы и best practices - Современные фреймворки автоматического ML - Тренды в области ИИ (LLM, мультимодальные модели) Уметь: Проектировать end-to-end ML пайплайны - Применять автоматизированный feature engineering - Оценивать бизнес-ценность ML-решений - Адаптировать state-of-the-art подходы под задачи - Владение MLflow, Kubeflow - Создание воспроизводимых ML-экспериментов - Автоматизация пайплайнов предобработки - A/B тестирование ML-моделей
PL-1	**Способен применять язык программирования Python для решения задач в области ИИ**
PL-1.3	Разрабатывает и поддерживает системы обработки больших данных различной степени сложности. Способен строить архитектуру вычислений с использованием cloud-native инструментов, в том числе бессерверных решений (Yandex Cloud Functions): Знает: архитектурные паттерны big data систем - Принципы serverless computing - Cloud-native подходы (контейнеризация, orchestration) - Асинхронное программирование в Python - Мониторинг и отладка распределенных систем. Умеет: Проектировать масштабируемые data-приложения - Использовать бессерверные архитектуры для ETL - Оптимизировать производительность Python-кода - Интегрировать различные cloud-сервисы. Владеет – навыками разработки на PySpark, Dask, Ray - Созданием и деплем cloud functions - Контейнеризацией приложений (Docker) - Настройкой автоматического масштабирования - Оптимизация costs в cloud-среде.

Содержание и структура дисциплины:

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.
Разделы дисциплины, изучаемые в _7_ семестре *(очная форма)*

№	Наименование разделов (тем)	Количество часов			
		Всего	Аудиторная работа		Внеаудиторная работа
			Л	ЛР	СРС
1	2	3	4	6	7
1.	Введение в Big Data. Проблематика и базовые концепции.	6	2	2	2
2.	**Распределенная файловая система HDFS и объектное хранение.** Hadoop и экосистема	6	2	2	2
3.	**Модели вычислений. MapReduce и его эволюция.**	6	2	2	2
4.	**Введение в Apache Spark. Архитектура и RDD.**	6	2	2	2
5.	**Spark SQL и DataFrames.**	6	2	2	2
6.	**Оптимизация в Spark.**	6	2	2	2
7.	**Работа с Spark в облаке и кластерном режиме.**	12	4	4	4
8.	**Потоковая обработка данных со Structured Streaming.**	8	2	2	4
9.	**ETL-пайпланы на Spark. Best Practices..**	6	2	2	2
10.	**OLAP vs OLTP. Колоночные базы данных.**	6	2	2	2
11.	**Глубокое погружение в ClickHouse. Движки таблиц и партиционирование.**	6	2	2	2
12.	**Оптимизация запросов в ClickHouse.**	6	2	2	2
13.	**Интеграция Spark и ClickHouse.**	6	2	2	2
14.	**Инструменты оркестрации данных. Apache Airflow.**	6	2	2	2
15	**Архитектура Big Data-решений на практике: кейсы индустриальных партнеров**	12	4	4	4
ИТОГО по разделам дисциплины		**104**	**34**	**34**	**36**
Контроль самостоятельной работы (КСР)		**4**			
Промежуточная аттестация (ИКР)		**0,3**			
Подготовка к текущему контролю		**35.7**			
Общая трудоемкость по дисциплине		**144**			

Примечание: Л – лекции, КСР – контрольные и самостоятельные работы, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

Курсовые проекты или работы.
Не предусмотрены учебным планом

Вид аттестации: ЛР, проект по кейсам индустриальных партнеров, экзамен.

Автор Приходько Т.А. – кандидат технических наук, доцент кафедры вычислительных технологий