Аннотация к рабочей программе дисциплины
Б1.О.15.04 Инжиниринг больших данных

Направление подготовки: 09.03.01

Квалификация выпускника: Квалификация выпускника: бакалавр

Цель освоения дисциплины: Целью изучения дисциплины "Инжиниринг больших данных" является формирование у студентов теоретических и практических знаний о методах сбора, извлечения и обработки данных, методах построения систем хранения данных, обучении технологии функционирования платформ распределенной обработки больших наборов данных.

Объем дисциплины: 4 зачетных единицы, всего 144 часа.

Семестр: 6

Краткое содержание основных разделов дисциплины:

№ п/п раздела	Краткое содержание разделов дисциплины
1	Раздел 1. Работа с данными в Python. Тема 1.1. Библиотеки для работы с данными в различных форматах в Python: файлы CSV, JSON, HTML. Работа с базами данных в Python. Тема 1.2. Работа с изображениями, видео и звуковыми файлами. Форматы хранения больших данных и работа с ними: Parquet, Avro. Графы знаний.
2	Раздел 2 Подготовка данных для систем машинного обучения. Тема 2.1. Сбор данных и формирование набора данных для систем машинного обучения. Загрузка данных из интернет и социальных сетей. Тема 2.2. Методы очистки и подготовки данных. Очистка и подготовка данных на Python. Разметка данных. Тема 2.3. Общедоступные платформы для хранения данных. Подход Data-Centric AI.
3	Раздел 3 Параллельная и распределенная обработка данных. Тема 3.1. Архитектура центров обработки данных, кластеры для параллельных и распределенных вычислений. Экосистема для распределенного хранения и обработки больших объемов данных: Apache Hadoop, Распределенная файловая система HDFS. Парадигма MapReduce. Тема 3.2. Решение задач с помощью MapReduce. Алгоритмы на графах в MapReduce. Pig и Hive. NoSQL базы данных: HBase и Cassandra. YARN. MapReduce 2.0 Тема 3.3. Распределенная обработка данных в Apache Spark. Архитектура Apache Spark: Resilient Distributed Dataset (RDD), действия трансформации. Работа с данными с использованием Spark DataFrame. Источники данных для Spark DataFrame. Обработка данных в Spark DataFrame. Использование SQL в Spark DataFrame

Форма промежуточной аттестации: экзамен