1. Наименование дисциплины

«Обработка статистических потоковых больших данных».

2. Место дисциплины в структуре образовательной программы

Дисциплина «Обработка статистических потоковых больших данных» является дисциплиной Модуля общепрофессиональных дисциплин направления подготовки 09.04.03-Прикладная информатика, направленности программы магистратуры: Управление большими данными.

Изучение дисциплины «Обработка статистических потоковых больших данных» базируется на знаниях, полученных в рамках дисциплин: «Нереляционные базы данных», «Проектирование информационных систем», «Современные компьютерные технологии машинного обучения».

3. Содержание дисциплины

Тема 1. Подходы к обработке больших данных

Основные понятия. Подходы к обработке больших данных. Экосистема Hadoop. Структура и принципы построения кластерных систем. Развертывание кластера. Прием данных для пакетной и интерактивной обработки: прием данных изоблака или локальных данных. Форматы файлов. Работа с текстовыми файлами. Файлы JSON. Значения, разделенные запятыми, и значения, разделенные табуляцией. SequenceFile. Объектные файлы. Форматы Hadoop для ввода и вывода. Сжатие файлов.

Тема 2. Обработка данных в Apache Spark

Управление Hadoop и Spark: создание и настройка кластера. Загрузка Spark. Введение в командные оболочки Spark. Введение в основные понятия Spark. Автономные приложения. Инициализация SparkContext. Отладка заданий Hadoop и Spark. Прием данных в Apache Spark. Программирование операций с RDD. Основы RDD. Создание RDD. Операции с RDD. Преобразования и действия. Отложенные вычисления. Передача функций в Spark.

Тема 3. SparkSQL

Работа с парами ключ/значение в Spark. Создание наборов пар. Преобразование наборов пар. Агрегирование, группировка, соединение, группировка. Управление распределением данных. Аккумуляторы. Широковещательные переменные. Работа с разделами по отдельности. Настройка Spark с помощью SparkConf. Включение SparkSQL в приложения. Использование SparkSQL в приложениях. Инициализация SparkSQL. Набор данных SchemaRDD. Кэширование. Загрузка и сохранение данных. Производительность SparkSQL.

Тема 4. Обработка потоковых данных

Архитектуры и абстракции потоковой обработки больших данных. Spark Streaming. Преобразования без сохранения состояния. Преобразования с сохранением состояния. Операции вывода. Источники исходных данных. Основные и дополнительные источники данных. Отказоустойчивость рабочих узлов и приемников. Проблемы производительности. Интервал пакетирования и протяженность окна. Степень параллелизма. Сборка мусора и использование памяти.