**МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ**

**ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ**

**«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»**
**(РУТ (МИИТ)**

Рабочая программа дисциплины (модуля),
как компонент образовательной программы
высшего образования - программа бакалавриата
по направлению подготовки
09.03.01 Информатика и вычислительная техника,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

# РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

## Методы анализа и обработки больших данных

Направление подготовки: 09.03.01 Информатика и вычислительная техника

Направленность (профиль): IT-сервисы и технологии обработки данных на транспорте

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

Целью освоение дисциплины (модуля) является приобретения практическиз навыков и умений, позволяющих создавать высокопроизводительные реализации известных методов вычислительной математики, анализа и обработки больших данных.

Задачами освоения дисциплины (модуля) являются:

освоение базовых знаний в области архитектуры современных многопроцессорных вычислительных систем параллельной обработки информации,

освоений технологий организации параллельных вычислений на многопроцессорных вычислительных комплексах с распределенной или общей оперативной памятью.

2. Планируемые результаты обучения по дисциплине (модулю).

Перечень формируемых результатов освоения образовательной программы (компетенций) в результате обучения по дисциплине (модулю):

**ПК-1** - Способен анализировать большие данные с использованием существующей в организации методологической и технологической инфраструктуры;

**ПК-7** - Способен к организации процессов разработки программного обеспечения .

Обучение по дисциплине (модулю) предполагает, что по его результатам обучающийся будет:

**Знать:**

программное обеспечение, необходимое для работы с большими данными,

необходимые элементы инфраструктуры для обработки больших данных.

**Уметь:**

Обрабатывать данные в Hadoop

Работать с распределенными файловыми системами в Hadoop

Настраивать кластеры в Hadoop

Составлять спецификации оборудования для различных задач

**Владеть:**

Инструментами настройки сетей

Знаниями о высокопроизводительных вычислениях

Навыками работы со статистическими параметрами вычислений
Технологиями параллельной обработки данных
Навыками работы с vCenter Server
Инструментами работы с Hadoop
Техниками настройки кластеров в Hadoop

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 3 з.е. (108 академических часа(ов).

3.2. Объем дисциплины (модуля) в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

| Тип учебных занятий | Количество часов | |
|---|---|---|
| | Всего | Семестр №5 |
| Контактная работа при проведении учебных занятий (всего): | 64 | 64 |
| В том числе: | | |
| Занятия лекционного типа | 32 | 32 |
| Занятия семинарского типа | 32 | 32 |

3.3. Объем дисциплины (модуля) в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации составляет 44 академических часа (ов).

3.4. При обучении по индивидуальному учебному плану, в том числе при ускоренном обучении, объем дисциплины (модуля) может быть реализован полностью в форме самостоятельной работы обучающихся, а также в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении промежуточной аттестации.

4. Содержание дисциплины (модуля).

4.1. Занятия лекционного типа.

| № п/п | Тематика лекционных занятий / краткое содержание |
|---|---|
| 1 | Тема 1. Большие данные. История развития методов анализа и обработки<br>Рассматриваемые вопросы:<br>- эволюция понятия больших данных<br>- хранилища данных<br>- история развития методов обработки больших данных<br>- отличия методов анализа от методов обработки больших данных |
| 2 | Тема 2. Жизненный цикл данных и метаданные<br>Рассматриваемые вопросы:<br>- элементы жизненного цикла и методы управления жизненным циклом<br>- метаданные |
| 3 | Тема 3. Архитектура систем обработки больших данных<br>Рассматриваемые вопросы:<br>- прием данных<br>- сбор данных<br>- анализ данных<br>- представления результатов<br>- витрины данных |
| 4 | Тема 4. Введение в понятия высокопроизводительных вычислений.<br>Рассматриваемые вопросы:<br>-Важность проблематики параллельных вычислений<br>-Пути достижения параллелизма. Векторная и конвейерная обработка данных.<br>- Многопроцессорная и многомашинная, параллельная обработка данных.<br>- Закон Мура, сдерживающие факторы наращивания количества транзисторов на кристалле и частоты процессоров. Привлекательность подхода параллельной обработки данных |
| 5 | Тема 5. Основные направления развития высокопроизводительных компьютеров.<br>Рассматриваемые вопросы:<br>-Сдерживающие факторы повсеместного внедрения параллельных вычислений<br>-Ведомственные, национальные и другие программы, направленные на развитие параллельных вычислений в России.<br>-Необходимость изучения дисциплины параллельного программирования. Перечень критических задач, решение которых без использования параллельных вычислений затруднено или вовсе невозможно. |
| 6 | Тема 6. Задачи параллельной обработки данных<br>Рассматриваемые вопросы:<br>- программные структуры алгоритмов параллельной обработки данных<br>- инструменты параллельной обработки данных<br>- конвейерно-параллельной обработки интегрированных потоков данных |
| 7 | Тема 7. Классификация вычислительных сетей<br>Рассматриваемые вопросы:<br>-Системы с распределенной, общей памятью, примеры систем.<br>-Массивно-параллельные системы (MPP). Симметричные мультипроцессорные системы (SMP). - Параллельные векторные системы (PVP).<br>-Системы с неоднородным доступом к памяти (Numa)<br>-Компьютерные кластеры – специализированные и полнофункциональные. История возникновения компьютерных кластеров–проект Beowulf. Метакомпьютинг. Классификация Флинна, Шора и т.д. Организация межпроцессорных связей – коммуникационные топологии.<br>-Примеры сетевых решений для создания кластерных систем |
| 8 | Тема 8. Основные принципы организации параллельной обработки данных: модели, методы и технологии параллельного программирования<br>Рассматриваемые вопросы: |

| № п/п | Тематика лекционных занятий / краткое содержание |
|---|---|
| | -Функциональный параллелизм, параллелизм по данным.<br>-Парадигма master-slave. Парадигма SPMD. Парадигма конвейеризации. Парадигма «разделяй и властвуй». Спекулятивный параллелизм. Важность выбора технологии для реализации алгоритма<br>-Модель обмена сообщениями – MPI.<br>-Модель общей памяти – OpenMP. Концепция виртуальной, разделяемой памяти – Linda.<br>Российские разработки – T-система, система DVM. Проблемы создания средства автоматического распараллеливания программ |
| 9 | Тема 9. Параллельное программирование с использованием интерфейса передачи сообщений MPI<br>Рассматриваемые вопросы:<br>-Библиотека MPI.<br>- Модель SIMD. Инициализация и завершение MPI-приложения.<br>Точечные обмены данными между процессами MPI-программы. Режимы буферизации.<br>Проблема deadlock'ов. Коллективные взаимодействия процессов в MPI. Управление группами и коммуникаторами в MPI |
| 10 | Тема 10. Параллельное программирование на системах с общей памятью (OpenMP)<br>Рассматриваемые вопросы:<br>-Введение в OpenMP<br>-Стандарты программирования для систем с разделяемой памятью.<br>-Создание многопоточных приложений. Использование многопоточности при программировании для многоядерных платформ.<br>-Синхронизация данных между ветвями в параллельной программе. Директивы языка OpenMP |
| 11 | Тема 11. Параллельное программирование многоядерных GPU. Кластеры из GPU и суперкомпьютеры на гибридной схеме<br>Рассматриваемые вопросы:<br>-Существующие многоядерные системы.<br>-GPU.<br>-Использование OpenMP и MPI технологий совместно с CUDA.<br>-Степень параллелизма численного алгоритма. Закон Амдала. Параллельный алгоритм решения СЛАУ |
| 12 | Тема 12. Программные платформы и системы для больших данных<br>Рассматриваемые вопросы:<br>- системы управления потоками данных<br>- системы хранения больших данных<br>- платформы больших данных<br>- обработка больших данных в реальном времени |

## 4.2. Занятия семинарского типа.

### Практические занятия

| № п/п | Тематика практических занятий/краткое содержание |
|---|---|
| 1 | Тема 1. Изучение технологий Hadoop и MapReduce<br>Рассматриваемые вопросы:<br>- Hadoop<br>- MapReduce |
| 2 | Тема 2. Анализ больших массивов данных инструментами python<br>Рассматриваемые вопросы: |

| № п/п | Тематика практических занятий/краткое содержание |
|---|---|
| | - библиотеки анализа и загрузки больших данных<br>- фреймворки Python с параллельной обработкой данных |
| 3 | Тема 3. Высокопроизводительные вычисления с фреймворком Apache Spark<br>Рассматриваемые вопросы:<br>- знакомство фреймворком<br>- знакомство с архитектурой системы<br>- реализации концепции MapReduce<br>- разработки программ с использованием Apache Spark |
| 4 | Тема 4. Введение в высокопроизводительные серверы на python<br>Рассматриваемые вопросы:<br>- работа с распределенными системами<br>- создание кода для работы на GRU |
| 5 | Тема 5. Параллельное программирование. Параллельная обработка данных<br>Рассматриваемые вопросы:<br>1. Параллельное программирование<br>2. Параллельная обработка данных |
| 6 | Тема 6. Работа с vCenter Server<br>Рассматриваемые вопросы:<br>1.Понятие vCenter<br>2. Установка и развертывание vCenter |
| 7 | Тема 7. Развертывание среды для обработки данных при помощи Hadoop<br>Рассматриваемые вопросы:<br>1. Обработка данных в Hadoop<br>2. Распределенная файловая система Hadoop<br>3. Разработка приложений MapReduce<br>4. Настройка кластера Hadoop |
| 8 | Тема 8. Составление спецификаций оборудования для работы с высокопроизводительными вычислениями<br>Рассматриваемые вопросы:<br>1. Составление спецификаций оборудования для работы с высокопроизводительными вычислениями |

## 4.3. Самостоятельная работа обучающихся.

| № п/п | Вид самостоятельной работы |
|---|---|
| 1 | Работа с учебной литературой |
| 2 | Участие в онлайн-конференциях и мастер-классах |
| 3 | Поиск алгоритмов обработки данных в открытых источниках |
| 4 | Подготовка к промежуточной аттестации. |
| 5 | Подготовка к текущему контролю. |

5. Перечень изданий, которые рекомендуется использовать при освоении дисциплины (модуля).

| № п/п | Библиографическое описание | Место доступа |
|---|---|---|
| 1 | Железнов, М. М. Методы и технологии обработки больших данных : учебно-методическое пособие / М. М. Железнов. — Москва : МИСИ – МГСУ, 2020. — 46 с. — ISBN 978-5-7264-2193-3. | https://e.lanbook.com/book/145102 |
| 2 | Ланских, Ю. В. Введение в большие данные : учебное пособие / Ю. В. Ланских, В. Г. Ланских, К. В. Родионов. — Киров : ВятГУ, 2023. — 172 с. | https://e.lanbook.com/book/408566 |

6. Перечень современных профессиональных баз данных и информационных справочных систем, которые могут использоваться при освоении дисциплины (модуля).

https://habr.com/ru - база знаний в виде статей, обзоров

https://journal.tinkoff.ru/short/ai-for-all/ - база данных нейронных сетей

https://vc.ru/services/916617-luchshie-neyroseti-bolshaya-podborka-iz-top-200-ii-generatorov-po-kategoriyam - база данных нейронных сетей

https://github.com/abalmumcu/bert-rest-api - профессиональная платформа для командой работы над проектов (нейронная сеть bert)

http://library.miit.ru/ - электронно-библиотечная система Научно-технической библиотеки МИИТ

https://proglib.io/p/raspoznavanie-obektov-s-pomoshchyu-yolo-v3-na-tensorflow-2-0-2020-11-08 - профессиональная библиотека программистов

https://yandex.cloud/ru/blog/posts/2022/12/andrey-berger-and-yandex-cloud?utm_referrer=https%3A%2F%2Fyandex.ru%2F – библиотека профессиональных статей разработчиков Яндекс

https://yandex.cloud/ru/blog - библиотека профессиональных статей разработчиков Яндекс

https://tproger.ru/translations/opencv-python-guide - библиотека основных команд OpenCV

7. Перечень лицензионного и свободно распространяемого программного обеспечения, в том числе отечественного производства, необходимого для освоения дисциплины (модуля).

Microsoft Office 2010

VMware Workstation

8. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю).

1 учебный класс
Компьютер преподавателя
Компьютеры студентов
Монитор
Проектор
Экран для проектора
Маркерная доска

9. Форма промежуточной аттестации:

Зачет в 5 семестре.

10. Оценочные материалы.

Оценочные материалы, применяемые при проведении промежуточной аттестации, разрабатываются в соответствии с локальным нормативным актом РУТ (МИИТ).

Авторы:

руководитель образовательной программы О.Б. Проневич

Согласовано:

Директор Д.В. Паринов

Руководитель образовательной программы О.Б. Проневич

Председатель учебно-методической комиссии Д.В. Паринов