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 - программы 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 5 з.е. (180 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	16	16
Занятия семинарского типа	48	48

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 116 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 Рассматриваемые вопросы: - эволюция понятия больших данных - хранилища данных - история развития методов обработки больших данных - отличия методов анализа от методов обработки больших данных
2	Тема 2. Жизненный цикл данных и метаданные Рассматриваемые вопросы: - элементы жизненного цикла и методы управления жизненным циклом - метаданные
3	Тема 3. Архитектура систем обработки больших данных Рассматриваемые вопросы: - прием данных - сбор данных - анализ данных - представления результатов - витрины данных
4	Тема 4. Введение в понятия высокопроизводительных вычислений. Рассматриваемые вопросы: -Важность проблематики параллельных вычислений -Пути достижения параллелизма. Векторная и конвейерная обработка данных. - Многопроцессорная и многомашинная, параллельная обработка данных. - 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. Рассматриваемые вопросы: -Сдерживающие факторы повсеместного внедрения параллельных вычислений -Ведомственные, национальные и другие программы, направленные на развитие параллельных вычислений в России. -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 Рассматриваемые вопросы: - программные структуры алгоритмов параллельной обработки данных - инструменты параллельной обработки данных - конвейерно-параллельной обработки интегрированных потоков данных
7	Тема 7. Классификация вычислительных сетей Рассматриваемые вопросы: -Системы с распределенной, общей памятью, примеры систем. -Массивно-параллельные системы (MPP). Симметричные мультипроцессорные системы (SMP). - Параллельные векторные системы (PVP). -Системы с неоднородным доступом к памяти (Numa) -Компьютерные кластеры – специализированные и полнофункциональные. История возникновения компьютерных кластеров–проект Beowulf. Метакомпьютинг. Классификация Флинна, Шора и т.д. Организация межпроцессорных связей – коммуникационные топологии. -Примеры сетевых решений для создания кластерных систем
8	Тема 8. Основные принципы организации параллельной обработки данных: модели, методы и технологии параллельного программирования Рассматриваемые вопросы:

№ п/п	Тематика лекционных занятий / краткое содержание
	-Функциональный параллелизм, параллелизм по данным. -Парадигма master-slave. Парадигма SPMD. Парадигма конвейеризации. Парадигма «разделяй и властвуй». Спекулятивный параллелизм. Важность выбора технологии для реализации алгоритма -Модель обмена сообщениями – MPI. -Модель общей памяти – OpenMP. Концепция виртуальной, разделяемой памяти – Linda. Российские разработки – T-система, система DVM. Проблемы создания средства автоматического распараллеливания программ
9	Тема 9. Параллельное программирование с использованием интерфейса передачи сообщений MPI Рассматриваемые вопросы: -Библиотека MPI. - Модель SIMD. Инициализация и завершение MPI-приложения. Точечные обмены данными между процессами MPI-программы. Режимы буферизации. 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) Рассматриваемые вопросы: -Введение в OpenMP -Стандарты программирования для систем с разделяемой памятью. -Создание многопоточных приложений. Использование многопоточности при программировании для многоядерных платформ. -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 Рассматриваемые вопросы: -Существующие многоядерные системы. -GPU. -Использование OpenMP и MPI технологий совместно с CUDA. -Степень параллелизма численного алгоритма. Закон Амдала. Параллельный алгоритм решения СЛАУ
12	Тема 12. Программные платформы и системы для больших данных Рассматриваемые вопросы: - системы управления потоками данных - системы хранения больших данных - платформы больших данных - 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 Рассматриваемые вопросы: - Hadoop - MapReduce
2	Тема 2. Анализ больших массивов данных инструментами python Рассматриваемые вопросы:

№ п/п	Тематика практических занятий/краткое содержание
	- библиотеки анализа и загрузки больших данных - фреймворки Python с параллельной обработкой данных
3	Тема 3. Высокопроизводительные вычисления с фреймворком Apache Spark Рассматриваемые вопросы: - знакомство фреймворком - знакомство с архитектурой системы - реализации концепции MapReduce - разработки программ с использованием Apache Spark
4	Тема 4. Введение в высокопроизводительные серверы на python Рассматриваемые вопросы: - работа с распределенными системами - создание кода для работы на GRU
5	Тема 5. Параллельное программирование. Параллельная обработка данных Рассматриваемые вопросы: 1. Параллельное программирование 2. Параллельная обработка данных
6	Тема 6. Работа с vCenter Server Рассматриваемые вопросы: 1.Понятие vCenter 2. Установка и развертывание vCenter
7	Тема 7. Развертывание среды для обработки данных при помощи Hadoop Рассматриваемые вопросы: 1. Обработка данных в Hadoop 2. Распределенная файловая система Hadoop 3. Разработка приложений MapReduce 4. Настройка кластера Hadoop
8	Тема 8. Составление спецификаций оборудования для работы с высокопроизводительными вычислениями Рассматриваемые вопросы: 1. Составление спецификаций оборудования для работы с высокопроизводительными вычислениями
9	Тема 9. Работа с базами данных NoSQL Рассматриваемые вопросы: - знакомство с базами данных NoSQL - сравнение SQL и NoSQL - работа с MongoDB и Cassandra
10	Тема 10. Машинное обучение на больших данных Рассматриваемые вопросы: - основы машинного обучения - применение машинного обучения для анализа больших данных - использование библиотек Scikit-learn и TensorFlow
11	Тема 11. Работа с облачными платформами для обработки данных Рассматриваемые вопросы: - знакомство с облачными платформами (AWS, Google Cloud, Azure) - развертывание кластеров в облаке - обработка данных в облачных средах

№ п/п	Тематика практических занятий/краткое содержание
12	Тема 12. Оптимизация производительности в распределенных Рассматриваемые вопросы: - методы оптимизации производительности - анализ узких мест в распределенных системах - инструменты мониторинга и анализа производительности
13	Тема 13. Работа с потоковыми данными Рассматриваемые вопросы: - обработка потоковых данных в реальном времени - использование Apache Kafka и Apache Flink - разработка приложений для обработки потоковых данных
14	Тема 14. Визуализация больших данных Рассматриваемые вопросы: - инструменты визуализации данных (Tableau, Power BI, Matplotlib) - создание интерактивных дашбордов - визуализация данных в реальном времени
15	Тема 15. Работа с графовыми базами данных Рассматриваемые вопросы: - знакомство с графовыми базами данных (Neo4j, ArangoDB) - анализ графовых данных - применение графовых баз данных в транспортных системах
16	Тема 16. Разработка и тестирование приложений для обработки Рассматриваемые вопросы: - разработка приложений для обработки больших данных - тестирование производительности и отказоустойчивости - оптимизация кода для работы с большими объемами данных

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	Габдуллин, Н. М. Развитие человеческого капитала и цифровой экономики в регионах России: факторный и кластерный анализ :	https://e.lanbook.com/book/173018

	монография / Н. М. Габдуллин. — Казань : КФУ, 2019. — 268 с. — ISBN 978-5-00130-291-9	
2	Глассner, Э. Глубокое обучение без математики. Том 2. Практика : руководство / Э. Глассner ; перевод с английского В. А. Яроцкого. — Москва : ДМК Пресс, 2020. — 610 с. — ISBN 978-5-97060-767-1	https://e.lanbook.com/book/131710
3	Кук, Д. Машинное обучение с использованием библиотеки H2O / Д. Кук ; перевод с английского А. Б. Огурцова. — Москва : ДМК Пресс, 2018. — 250 с. — ISBN 978-5-97060-508-0	https://e.lanbook.com/book/97353
4	Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флах. — Москва : ДМК Пресс, 2015. — 400 с. — ISBN 978-5-97060-273-7	https://e.lanbook.com/book/69955
5	Шалев-Шварц, Ш. Идеи машинного обучения : учебное пособие / Ш. Шалев-Шварц, Бен-ДавидШ. ; перевод с английского А. А. Слинкина. — Москва : ДМК Пресс, 2019. — 436 с. — ISBN 978-5-97060-673-5	https://e.lanbook.com/book/131686
6	Гудфеллоу, Я. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль ; перевод с английского А. А. Слинкина. — 2-е изд. — Москва : ДМК Пресс, 2018. — 652 с. — ISBN 978-5-97060-618-6	https://e.lanbook.com/book/107901

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:

старший преподаватель кафедры
«Цифровые технологии управления
транспортными процессами» И.В. Зенковский

Согласовано:

Директор Б.В. Игольников

Руководитель образовательной
программы О.Б. Проневич

Председатель учебно-методической
комиссии Д.В. Паринов