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).

Дисциплина Методы анализа и обработки больших данных, программное обеспечение: Библиотеки Python для Data Science: NumPy, Matplotlib, Scikit-learn предназначена для того, чтобы дать знания, умения и основные навыки, позволяющие создавать высокопроизводительные реализации известных методов вычислительной математики, анализа и обработки данных. Целью освоения дисциплины является – освоение базовых знаний в области архитектуры современных многопроцессорных вычислительных систем параллельной обработки информации, 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):

ОПК-3 - Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности;

ОПК-4 - Способен участвовать в разработке стандартов, норм и правил, а также технической документации, связанной с профессиональной деятельностью;

ОПК-9 - Способен осваивать методики использования программных средств для решения практических задач;

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

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:

Строить архитектуру сетей

Прописывать документацию под сети

Настраивать сети

Настраивать статическую и динамическую маршрутизацию

Настраивать VPN

Разбираться в типах микропроцессорных ВС

Составлять алгоритмы параллельного программирования

Декомпозировать задачи, для которых необходимо применение технологии DevOpsУстанавливать и настраивать vCenter Server

Обрабатывать данные в Hadoop

Работать с распределенными файловыми системами в Hadoop

Разрабатывать приложения MapReduce

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

Инструментами разработки приложений MapReduce

Техниками составления спецификации оборудования для различных задач

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 4 з.е. (144 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	Сем. №5
Контактная работа при проведении учебных занятий (всего):	80	80
В том числе:		
Занятия лекционного типа	32	32
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 64 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	Введение в понятия высокопроизводительных вычислений. Основные направления развития высокопроизводительных компьютеров. 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 тран-зисторов на кристалле и частоты процес-соров. Привлекательность подхода па-раллельной обработки данных -Сдерживающие факторы повсеместного внедрения параллельных вычислений -Ведомственные, национальные и другие программы, направленные на развитие параллельных вычислений в России. Необходимость изучения дисциплины параллельного программирования. Пере-чень критических задач, решение кото-рых без использования параллельных вычислений затруднено или вовсе невоз-можно.
2	Классификация микропроцессорных ВС Рассматриваемые вопросы: -Системы с распределенной, общей памятью, примеры систем. -Массивно-параллельные системы (MPP). Симметричные мультипроцессорные системы (SMP). -Параллельные векторные системы (PVP).

№ п/п	Тематика лекционных занятий / краткое содержание
	-Системы с неоднородным доступом к памяти (Numa) -Компьютерные кластеры – специализированные и полнофункцио-нальные. История возникновения компь-ютерных кластеров–проект Beowulf. Мета-компьютинг. Классификация Флинна, Шора и т.д. Организация межпроцессор-ных связей – коммуникационные топологии. -Примеры сетевых решений для со-здания кластерных систем
3	Основные принципы организации параллельной обработки данных: модели, методы и технологии параллельного программирования Рассматриваемые вопросы: -Функциональный параллелизм, паралле-лизм по данным. -Парадигма master-slave. Парадигма SPMD. Парадигма конвейеризации. Парадигма «разделяй и властвуй». Спекулятивный параллелизм. Важность выбора технологии для реализации алго-ритма -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
4	Параллельное программирование с использованием интерфейса передачи сообщений MPI Рассматриваемые вопросы: -Библиотека MPI. - Модель SIMD. Инициализация и завершение MPI-приложения. Точечные обмены данными между про-цессами MPI-программы. Режимы буферизации. Проблема deadlock'ов. Коллективные взаимодействия процессов в MPI. Управление группами и коммуникаторами в MPI
5	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 про-грамме. Директивы языка OpenMP
6	Параллельное программирование многоядерных GPU. Кластеры из GPU и супер-компьютеры 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 реше-ния СЛАУ

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Практические задания на основе библиотек Python (NumPy, Matplotlib, Scikit-learn) Рассматриваемые вопросы: 1. Планирование архитектуры 2. Подготовка документации для архитектуры 3. Подключение к оборудованию cisco и настройка сети

№ п/п	Тематика практических занятий/краткое содержание
	4. Статическая маршрутизация 5. Динамическая маршрутизация и VPN
2	Высокопроизводительные вычисления. Классификация микропроцессорных ВС Рассматриваемые вопросы: 1. Высокопроизводительные вычисления 2. Классификация ВС
3	Параллельное программирование. Параллельная обработка данных Рассматриваемые вопросы: 1. Параллельное программирование 2. Параллельная обработка данных
4	Работа с vCenter Server Рассматриваемые вопросы: 1.Понятие vCenter 2. Установка и развертывание vCenter
5	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
6	Составление спецификаций оборудования для работы с высокопроизводительными вычислениями Рассматриваемые вопросы: 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	Выполнение курсовой работы.
5	Подготовка к промежуточной аттестации.
6	Подготовка к текущему контролю.

4.4. Примерный перечень тем курсовых работ

1. Анализ данных с использованием алгоритмов кластеризации
2. Кластеризация данных с помощью нечетких отношений
3. Метрики, применяемые в Data mining
4. Основные стандарты Data mining

5. Направления использования эволюционных алгоритмов анализа данных

6. Анализ данных с использованием генетических алгоритмов

7. Применение методов Data mining для решения практических задач

8. Технология Knowledge Discovery in Databases (KDD)

9. Характеристики промышленных инструментальных средств Data mining

10. Использование реляционной модели построения хранилищ данных (ROLAP)

11. Использование многомерного подхода в построении хранилищ данных (MOLAP)

12. Использование гибридных (HOLAP) и виртуальных хранилищ данных

13. Технологии и методы оценки качества, очистки и предобработки анализируемых данных

14. Технология практического применения сэмплинга (sampling)

15. Сущность и направления использования аффинитивного анализа данных

16. Подходы к решению задач поиска ассоциативных правил

17. Анализ данных с использованием сети Кохонена (Kohonen network)

18. Анализ данных с использованием самоорганизующихся карт Кохонена (Self organizing map)

19. Технология анализа данных с применением регрессионных моделей

20. Технология построение и оценка значимости простой регрессионной модели

21. Характеристика алгоритмов построения деревьев решений

22. Подготовка управленческих решений на основе метода деревьев решений

23. Принципы построения и направления практического применения нейросетевых моделей

24. Подходы к анализу данных на базе ансамблей моделей

25. Применение моделей анализа временных рядов

26. Технологии обогащения данных

27. Технологии упрощения деревьев решений

28. Алгоритмы обучения нейронных сетей

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

№ п/п	Библиографическое описание	Место доступа
1	Глубокое обучение Гудфеллоу Я., Бенджио И., Курвилль А.	https://e.lanbook.com/book/107901
2	Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных Флах П.	https://e.lanbook.com/book/69955
3	Математические методы распознавания образов Местецкий Л.М	https://e.lanbook.com/book/100634
4	Габдуллин, Н. М. Развитие человеческого капитала и цифровой экономики в регионах России: факторный и кластерный анализ : монография / Н. М. Габдуллин. — Казань : КФУ, 2019. — 268 с. — ISBN 978-5-00130-291-9. — Текст : электронный // Лань : электронно-библиотечная система	https://e.lanbook.com/book/173018
5	Гитис, Л. Х. Статистическая классификация и кластерный анализ / Л. Х. Гитис. — Москва : Горная книга, 2003. — 157 с. — ISBN 5-7418-0010-6. — Текст : электронный // Лань : электронно-библиотечная система.	https://e.lanbook.com/book/3493
6	Гласснер, Э. Глубокое обучение без математики. Том 2. Практика : руководство / Э. Гласснер ; перевод с английского В. А. Яроцкого. — Москва : ДМК Пресс, 2020. — 610 с. — ISBN 978-5-97060-767-1. — Текст : электронный // Лань : электронно-библиотечная система	https://e.lanbook.com/book/131710
7	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. — Текст : электронный // Лань : электронно-библиотечная система.	https://e.lanbook.com/book/97353
8	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. — Текст : электронный // Лань : электронно-библиотечная система.	https://e.lanbook.com/book/69955

9	Целых, А. Н. Современные технологии противодействия финансовым преступлениям : учебное пособие / А. Н. Целых. — Ростов-на-Дону : ЮФУ, 2019. — 119 с. — ISBN 978-5-9275-3286-5. — Текст : электронный // Лань : электронно-библиотечная система	https://e.lanbook.com/book/141063
10	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. — Текст : электронный // Лань : электронно-библиотечная система.	https://e.lanbook.com/book/131686

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

https://habr.com/ru/

https://e.lanbook.com/

https://rusneb.ru/

https://www.vmware.com

Про HPE Synergy, часть I — Вступление [Электронный ресурс] URL: https://habr.com/ru/post/308224/

Про HPE Synergy, часть II – Шасси и сервера [Электронный ресурс] URL: https://habr.com/ru/post/310092/

Про HPE Synergy. Часть III – Дисковое хранилище D3940 и SAS-коммутаторы [Электронный ресурс] URL: https://habr.com/ru/post/310564/

Про HPE Synergy – часть IV. Наши сети [Электронный ресурс] URL: https://habr.com/ru/post/313240/

Про HPE Synergy – часть V. Управление [Электронный ресурс] URL: https://habr.com/ru/post/319430/

Обзор новой линейки систем хранения данных HP 3PAR [Электронный ресурс] URL: https://habr.com/ru/company/muk/blog/263469/

https://habr.com/ru/post/136056/,

https://linkmeup.ru/blog/1190/

Сети для самых маленьких. Часть 8-12 [Электронный ресурс] URL: https://linkmeup.ru/blog/1198/

Основы компьютерных сетей. [Электронный ресурс] URL: https://habr.com/ru/post/307252/

A Review of Supercomputer Performance Monitoring Systems [Электронный ресурс] URL: https://superfri.org/index.php/superfri/article/view/392

Высокопроизводительные вычислительные платформы: текущий статус и тенденции развития [Электронный ресурс] URL: https://en.num-meth.ru/index.php/journal/article/view/1160

Parallel structure of algorithms and training computational technology specialists [Электронный ресурс] URL: https://istina.msu.ru/publications/article/193994645/

About vCenter Server Installation and Setup [Электронный ресурс] URL: https://docs.vmware.com/en/VMware-vSphere/7.0/com.vmware.vcenter.install.doc/GUID-8DC3866D-5087-40A2-8067-1361A2AF95BD.html

Hadoop: The Definitive Guide [Электронный ресурс] URL: https://web.cs.dal.ca/~allen/HadoopDefinitiveGuide.pdf (на английском языке)

Hadoop: Подробное руководство [Электронный ресурс] URL: https://disk.yandex.ru/i/xpt3r337r01S-A (на русском языке)

https://h41370.www4.hpe.com/products/quickspecs/hppb_catalogs/hppb_installer.exe

Kubernetes Components [Электронный ресурс] URL: https://kubernetes.io/docs/concepts/overview/components/.

Руководство по Kubernetes, часть 2: создание кластера и работа с ним [Электронный ресурс] URL: https://habr.com/ru/company/ruvds/blog/438984/.

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 (столы, стулья - по 25 ед)

Компьютер преподавателя

Intel Core i7-9700 / Asus PRIME H310M-R R2.0 / 2x8GB / SSD 250Gb / DVDRW

Компьютеры студентов (24 ед)

Intel Core i9-9900 / B365M Pro4 / 2x16GB / SSD 512Gb

Монитор (25 ед)
Проектор Optoma W340UST
Экран для проектора
Маркерная доска

9. Форма промежуточной аттестации:

Курсовая работа в 5 семестре.
Экзамен в 5 семестре.

10. Оценочные материалы.

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

Авторы:

И.В. Зенковский

Согласовано:

Заместитель директора академии	Д.В. Паринов
Председатель учебно-методической комиссии	Д.В. Паринов