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)

Технологии хранения 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 освоения учебной дисциплины «Технологии хранения больших данных» является теоретическая и практическая подготовка студентов к работе с большими данными. Знания и компететенция, полученные в результате освоения дисциплины, помогут при сборе и анализе больших объемов структурированной или неструктурированной информации, при разработке моделей данных и получении новых знаний. Все это необходимо выпускнику, освоившему программу бакалавриата, для решения различных задач в области разработки корпоративных информационных систем и сервисов.

Задачи освоения дисциплины:

• приобретение студентами знаний о технологиях подготовки, хранения, обработки и анализа больших данных;

• применение статистических и математических методов для анализа больших объемов информации;

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-5 - Способен инсталлировать программное и аппаратное обеспечение для информационных и автоматизированных систем;

ОПК-6 - Способен разрабатывать бизнес-планы и технические задания на оснащение отделов, лабораторий, офисов компьютерным и сетевым оборудованием;

ОПК-8 - Способен разрабатывать алгоритмы и программы, пригодные для практического применения;

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:

методы анализа, хранения и обработки больших данных, интегрируемых в корпоративные информационные системы

Уметь:

Находить данные для анализа в открытых источниках

Загружать данные, проводить предварительную чистку данных для анализа

Проводить анализ данных в зависимости от их типа

Строить графики и проводить статистический анализ на их основе

Делать выводы, исходя из статистического анализа данных

Проводить MCA и CA анализ

Интерпретировать полученные результаты и проверять на основе их различные гипотезы

Проводить установку Apache Hadoop

Настраивать в минимальных требованиях Apache Hadoop

Тестировать Apache Hadoop

Форматировать файловую систему

Проводить установку Spark

Проводить установку Spark на Ubuntu

Настраивать в минимальных требованиях Spark

Устанавливать программы на Spark

Работать с библиотекой NumPy и SciPy

Проводить анализ различных функций с помощью библиотеки NumPy

Работать с матрицами

Проводить вычисления по цепочке

Собирать данные

Работать с библиотекой Pandas

Извлекать данные из датасета

Загружать датасет

Создавать новые столбцы и строки

Работать с имеющимися данными датасета и проводить различный анализ

Пользоваться библиотекой matplotlib

Проводить анализ данных на основе визуализации

Строить графики различного характера для анализа данных

Различать графические формы визуализации данных

Владеть:

Инструментами и командами языка R для анализа данных

Знаниями об определении типа диаграмм

Статистическими параметрами, коэффициентами и тестами
Инструментами MCA и CA анализа
Тремя режимами установки Apache Hadoop
Системой установки Spark
Инструментами библиотеки NumPy и SciPy
Инструментами библиотеки Pandas
Инструментами визуализации данных на основе библиотеки matplotlib

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. №3
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 80 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	Методы многомерного статистического анализа и анализа нечисловой информации 1.Факторный анализ. 2.Дискриминантный анализ. 3.Кластерный анализ. 4.Многомерное шкалирование. 5.Методы контроля качества.
2	Технологии хранения и обработки больших данных 1.Основные направления развития методов обработки и хранения данных. 2.Закон Мура. 3.Фреймворк Hadoop. 4.Проблема хранения неструктурированных данных. 5.Проблема преобразования данных. 6.Семантические анализаторы. 7.Само-обучающиеся автоматы.
3	Программирование обработки и загрузки больших данных 1. 9 языков для Big Data (R, Python, Julia, Java, Scala, MATLAB, Go, Kafka, Hadoop). Фреймворки (Hadoop, Spark, Storm). 2.Базы данных (Hive, Impala, Presto, Drill). 3.Аналитические платформы (Rapid Miner, IBM SPSS Modeler, KNIME, Qlik Analytics Platform, STATISTICA Data Miner, Informatica Intelligent Data Platform, World Programming System, Deductor, SAS Enterprise Miner). 4.Прочие инструменты (Zookeeper, Flume, IBM Watson Analytics, Dell EMC Analytic Insights Module, Windows Azure HDInsight, Microsoft Azure Machine Learning, Pentaho Data Integration, Teradata Aster Analytics, SAP BusinessObjects Predictive Analytics, Oracle Big Data Preparation).
4	Аналитика в больших данных 1.Аналитика Big Data — реалии и перспективы в России и мире. 2. Технологии и методы анализа, которые используются для анализа Big Data(Data Mining; краудсорсинг; смешение и интеграция данных; машинное обучение; искусственные нейронные сети; распознавание образов; прогнозная аналитика; имитационное моделирование; пространственный анализ; статистический анализ; визуализация аналитических данных).

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Методы многомерного статистического анализа и анализа нечисловой информации Анализ данных в R Методы анализа данных
2	Практические задания на основе набора утилит, библиотек и фреймворка Hadoop Установка виртуальной машины Установка Hadoop Минимальная настройка Apache Hadoop

№ п/п	Тематика практических занятий/краткое содержание
	Тестирование Hadoop. Установка Hadoop: Распределенный режим (локально).
3	Практические работы на основе фреймворка Apache Spark Установка Apache Spark Установка Apache Spark в Ubuntu
4	Программирование обработки и загрузки больших данных с использованием утилит и библиотек Python Библиотека NumPy Библиотека Pandas
5	Аналитика в больших данных Визуализация

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Технологии работы систем прогнозирования стоимости биржевых товаров на основе информации из СМИ
2	Применение технологий Big Data для повышения рейтинга игр на IMDb
3	Эпидемиологические исследования при помощи Big Data (Применение технологий Big Data в здравоохранении)
4	Цели и методы применения технологий Big Data для анализа социальных сетей
5	Анализ оптимальных авиамаршрутов по стране с помощью технологий Big Data
6	Принципы хранения больших данных в рекомендательной маркетинговой системе
7	Особенности и принципы работы аналитической платформы ClickHouse
8	Поиск корреляци рейтинга фильма с его жанром на Кинопоиске с помощью алгоритмов Big
9	Анализ применимости методов обработки данных чековых транзакций и их сравнение
10	Алгоритмы обработки данных для поиска проблемных станции по данным ЦФТО "РЖД"
11	Выполнение курсовой работы.
12	Подготовка к промежуточной аттестации.
13	Подготовка к текущему контролю.

4.4. Примерный перечень тем курсовых работ

Анализ, обработка, визуализация пассажиропотоков Московского транспортного узла (варианты по видам транспорта и направлениям)

1. Применение R для анализа данных

2. Методы анализа больших данных

3. Факторный анализ больших данных

4. Аналитика Big Data — реалии и перспективы в России и мире.

5. Технологии и методы анализа, которые используются для анализа Big Data

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

№ п/п	Библиографическое описание	Место доступа
1	J. Verzani, Using R for Introductory Statistics, Second Edition, Chapman & Hall/CRC The R Series, Taylor & Francis	books.google.ru/books?id=O86uAwAAQBAJ
2	B. Everitt, T. Hothorn, An introduction to applied multivariate analysis with R, Springer, New York, 2011	http://dx.doi.org/10.1007/978-1-4419-9650-3
3	Уайт Т. Hadoop: Подробное руководство. – СПб.: Питер, 2013. – 672 с.: ил. – (Серия «Бестселлеры O'Reilly»)	https://www.labirint.ru/books/396848/
4	Карау Х., Конвински Э. Изучаем Spark: молниеносный анализ данных. – М.: ДМК Пресс, 2015. – 304 с.: ил	https://books.google.ru/books?id=tc1SEAAAQBAJ
5	1. Shvachko, Konstantin. Apache Hadoop. The Scalability Update (англ.). — 2011. — Vol. 36, no. 3. — P. 7—13. — ISSN 1044-6397	http://home.apache.org/~shv/Publications.html
6	Дэвидсон-Пайлон К. Вероятностное программирование на Python: байесовский вывод и алгоритмы. – СПб.; Питер, 2019. – 256 с.: ил. – (Серия «Библиотека программиста»).	https://www.labirint.ru/books/702249/
7	Абдрахманов М. Devpractice Team. Pandas. Работа с данными. 2-е изд. – devpractice.ru. 2020. – 170 с.: ил	https://books.google.ru/books?id=tc1SEAAAQBAJ

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

Python для сетевых инженеров [сайт]

ЭБС Лань

Национальная электронная библиотека

The home of the U.S. Government’s open data

L. Torgo, Data Mining with R, learning with case studies, Chapman and Hall/CRC, 2010.

H. Georgakopoulos, Quantitative Trading with R: Understanding Mathematical and Compu- tational Tools from a Quant’s Perspective, Palgrave Macmillan, 2015.

Wickham, H., & Grolemund, G. (2016). R for Data Science?: Import, Tidy, Transform, Visualize, and Model Data (Vol. First edition). Sebastopol, CA: Reilly - O’Reilly Media

Роберт И., Кабаков - R в действии. Анализ и визуализация данных в программе R - Издательство "ДМК Пресс" - 2014 - 588с. - ISBN: 978-5-97060-077-1

How to Install Hadoop on Ubuntu 18.04 or 20.04

Apache Hadoop 2021.

Hadoop: Setting up a Single Node Cluster: [сайт]. – 2021.

Big Data School. – 2021.

Знакомство с Apache Spark

Apache Spark: [сайт]

apache / spark: [сайт]

Scala/ Programming language

3Apache Software Foundation Distribution Directory: [сайт]

Spark Standalone Mode: [сайт]. – 2021.

How To Install Apache Spark on Ubuntu: [сайт]. – 2021.

3Download Apache Spark: [сайт]. – 2021.

W1L1: Introduction to Big Data with Apache Spark: Spark Tutorial Lab: [сайт]. – 2021l

NumPy Reference: [сайт]. – 2021

NumPy для начинающих: [сайт]. – 2021

NumPy в Python. Часть 1: [сайт]. – 2021

Pandas: [сайт]. ¬– 2021

IO tools (text, CSV, HDF5, …): [сайт]

Pandas. Lessons: [сайт]. – 2021

Введение в анализ данных с помощью Pandas: [сайт]. – 2021

Matplotlib: [сайт]. – 2021

Дж. Вандер Плас. Python для сложных задач. Наука о данных и машинное обучение = Python Data Science Handbook: Essential Tools for Working with Data. — Питер, 2017. — 576 с. — ISBN 978-5-496-03068-7.

Toby Segaran. Programming Collective Intelligence: Building Smart Web 2.0 Applications. — O'Reilly Media, Inc., 2007. — 308 с. — ISBN 9780596529321. Имеется перевод: Тоби Сегаран. Программируем коллективный разум. — Символ-Плюс, 2009. — 368 с. — ISBN 5-93286-119-3.

Sandro Tosi. Matplotlib for Python Developers. — Packt Publishing, 2009. — 308 с. — ISBN 978-1847197900.

Shai Vaingast. Beginning Python Visualization: Crafting Visual Transformation Scripts. — Springer, 2009. — 384

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

Windows 7 и выше
Microsoft Office 2010
R consoler

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 3 семестре.
Экзамен в 3 семестре.

10. Оценочные материалы.

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

Авторы:

доцент, к.н. Академии "Высшая инженерная школа" Б.В. Игольников

Согласовано:

Заместитель директора академии Д.В. Паринов

Председатель учебно-методической комиссии Д.В. Паринов