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 магистратуры
по направлению подготовки
09.04.01 Информатика и вычислительная техника,
утвержденной первым проректором РУТ (МИИТ)
Тимониным В.С.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

Технологии больших данных

Направление подготовки: 09.04.01 Информатика и вычислительная техника

Направленность (профиль): Информационная аналитика и технология больших данных

Форма обучения: Очная

1. Общие сведения о дисциплине (модуле).

Целью изучения данного курса является получение студентами знаний о больших данных, ключевых принципах и технологиях их обработки, принципах и инструментах, используемых при работе с большими данными.

Задачи освоения дисциплины следующие:

- изучение подходов применяемых в работе с большими данными;
- ознакомление с инструментами, используемыми при работе с большими данными;
- освоение технологии MapReduce, инструментов HADOOP, HIVE и языка R.

Дисциплина предназначена для получения знаний и решения следующих профессио-нальных задач (в соответствии с видами деятельности):

Научно-исследовательская деятельность:

• разработка рабочих планов и программ проведения научных исследований и технических разработок, подготовка отдельных заданий для исполнителей;

• сбор, обработка, анализ и систематизация научно-технической информации по теме исследования, выбор методик и средств решения задачи;

• разработка математических моделей исследуемых процессов и изделий;

• разработка методик проектирования новых процессов и изделий;

• разработка методик автоматизации принятия решений;

• организация проведения экспериментов и испытаний, анализ их результатов;

• подготовка научно-технических отчетов, обзоров, публикаций по результатам выполненных исследований.

Проектная деятельность:

• подготовка заданий на разработку проектных решений;

• разработка проектов автоматизированных систем различного назначения, обоснование выбора аппаратно-программных средств автоматизации и информатизации предприятий и организаций;

• концептуальное проектирование сложных изделий, включая программные комплексы, с использованием средств автоматизации проектирования, передового опыта разработки конкурентноспособных изделий;

• выполнение проектов по созданию программ, баз данных и комплексов программ автоматизированных информационных систем;

• разработка и реализация проектов по интеграции информационных систем в соответствии с методиками и стандартами информационной поддержки изделий, включая методики и стандарты документооборота, интегрированной логистической поддержки, оценки качества программ и баз данных, электронного бизнеса;

• проведение технико-экономического и функционально-стоимостного анализа эффективности проектируемых систем;

• разработка методических и нормативных документов, технической документации, а также предложений и мероприятий по реализации разработанных проектов и программ.

Производственно-технологическая деятельность:

• проектирование и применение инструментальных средств реализации программно-аппаратных проектов;

• разработка методик реализации и сопровождения программных продуктов;

• разработка технических заданий на проектирование программного обеспечения для средств управления и технического оснащения промышленного производства и их реализация с помощью средств автоматизированного проекти-рования;

• тестирование программных продуктов и баз данных;

• выбор систем обеспечения экологической безопасности производства.

В результате изучения дисциплины студенты должны получить необходимые знания об ключевых принципах и технологиях обработки больших данных, принципах и инструментах, используемых при работе с большими данными.

2. Планируемые результаты обучения по дисциплине (модулю).

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

ОПК-4 - Способен применять на практике новые научные принципы и методы исследований;

ОПК-8 - Способен осуществлять эффективное управление разработкой программных средств и проектов.;

ПК-4 - Способность формировать технические задания и участвовать в разработке программных средств вычислительной техники;

ПК-6 - Владение существующими методами и алгоритмами решения задач цифровой обработки сигналов;

ПК-12 - Применение перспективных методов исследования и решения профессиональных задач на основе знания мировых тенденций развития вычислительной техники и информационных технологий.

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

Уметь:

- производить обработку больших данных с использованием HADOOP, HIVE, технологии MapReduce, языка R в среде RStudio.

Знать:

- технологии обработки больших данных, методы проведения исследования больших данных, методы разработки программного обеспечения на языке R; методы исследования и решения профессиональных задач; мировые тенденции развития вычислительной техники; знать перспективные тенденции развития информационных технологий.

Владеть:

- навыками применения технологии MapReduce, фрймворка HADOOP, языка R в обработке больших данных.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 6 з.е. (216 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):	34	34
В том числе:		
Занятия лекционного типа	18	18
Занятия семинарского типа	16	16

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 182 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	Понятие больших данных. Рассматриваемые вопросы: - определение больших данных; - основные принципы работы с большими данными; - горизонтальная масштабируемость; - отказоустойчивость; - локальность данных.
2	Технология MapReduce. Рассматриваемые вопросы: - модель распределённых вычислений, а также её реализации, используемые для параллельной обработки больших объёмов информации; - описание шага Map; - описание шага Reduce; - реализации MapReduce.
3	Apache Hadoop. Рассматриваемые вопросы: - файловая система HDFS; - движки: MapReduce, Spark, Tez; - реляционные БД: Hive, Impala, Shark, Spark SQL, Drill; - нереляционные БД: HBase; - форматы данных: Parquet, ORC, Thrift, Avro.
4	Hive. Рассматриваемые вопросы: - установка Hive; - объекты, которыми оперирует Hive: база данных, таблица, партиция (partition), бакет (bucket).
5	R — мультипарадигмальный интерпретируемый язык программирования для статистической обработки данных и работы с графикой.

№ п/п	Тематика лекционных занятий / краткое содержание
	Рассматриваемые вопросы: - начало работы и получение справочной информации, загрузка и установка R, работа в RStudio; - основы языка R; - ввод-вывод в языке R; - структуры данных в языке R; - преобразования данных с помощью языка R.
6	Линейная регрессия и дисперсионный анализ для больших данных, с использованием языка R. Рассматриваемые вопросы: - простая линейная регрессия; - выбор наиболее подходящих переменных регрессии; - поиск наиболее подходящего степенного преобразования (тест Бокса–Кокса); - формирование доверительных интервалов для коэффициентов регрессии; - обнаружение влиятельных наблюдений.
7	Анализ временных рядов для больших данных, с использованием языка R. Рассматриваемые вопросы: - представление данных временного ряда; - извлечение самых старых или самых последних наблюдений; - вычисление последовательных различий; - выполнение расчетов по временным рядам; - построение прогноза.

4.2. Занятия семинарского типа.

Лабораторные работы

№ п/п	Наименование лабораторных работ / краткое содержание
1	Установка и настройка Apache Hadoop. В результате выполнения лабораторной работы студент получает навык установки и настройки Apache Hadoop.
2	Работа с файловой системой HDFS. В результате выполнения лабораторной работы студент получает навык работы с распределённой файловой системой HDFS.
3	Обработка больших данных с использованием MapReduce. В результате выполнения лабораторной работы студент получает навык обработки больших данных с использованием MapReduce.
4	Работа с Hive. В результате выполнения лабораторной работы студент получает навыки установки Hive и обработки больших данных с использованием Hive.
5	Установка R и работа в RStudio. В результате выполнения лабораторной работы студент получает навыки установки R и работа в RStudio.
6	Линейная регрессия и дисперсионный анализ для больших данных, с использованием языка R. В результате выполнения лабораторной работы студент получает навыки применения линейной регрессии и дисперсионного анализа к большим данным, с использованием языка R.

№ п/п	Наименование лабораторных работ / краткое содержание
7	Анализ временных рядов для больших данных, с использованием языка R. В результате выполнения лабораторной работы студент получает навыки применения анализ временных рядов для больших данных, с использованием языка R.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Использование Spark в обработке больших данных. В результате работы на практическом занятии студент получает навыки работы со Spark при обработки больших данных.
2	Множественная линейная регрессия для больших данных, с использованием языка R. В результате работы на практическом занятии студент получает навыки применения множественной линейной регрессии для больших данных, с использованием языка R
3	Полиномиальная регрессия для больших данных, с использованием языка R. В результате работы на практическом занятии студент получает навыки применения полиномиальной регрессии для больших данных, с использованием языка R.
4	Построение функции автокорреляции на больших данных с использованием языка R. В результате работы на практическом занятии студент получает навыки построения функции автокорреляции на больших данных с использованием языка R.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Общие определения. Самостоятельное изучение теоретического материала раздела дисциплины. Источники: основная рекомендуемая литература .
2	Изучение работы со строками и датами в языке R.
3	Работа с вероятностью в языке R.
4	Сбор статистики для больших данных с использованием языка R. Изучение передовых статистических методов.
5	Подготовка к промежуточной аттестации.
6	Подготовка к текущему контролю.

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

№ п/п	Библиографическое описание	Место доступа
1	R. Книга рецептов. Проверенные	https://batrachos.com/sites/default/files/pictures/Books/Long_Titor_2020_R_Cookbook.pdf

	рецепты для статистики, анализа и визуализации Дж. Д. Лонг и Пол Титор, ДМК Пресс Москва, 2020, 510 с., ISBN 978-5-97060-835-7	
2	Документация по HADOOP.	https://hadoop.apache.org/
3	Григорьев Ю.А. Анализ свойств баз данных NoSQL. Научный журнал "Информатика и системы управления", 2013. № 2 (36). С. 003-013	https://elibrary.ru/item.asp?id=19060114 (дата обращения: 03.10.2022)
4	Труды Института системного программирования: Том 24, /Под ред. Академика РАН В.П. Иванникова/ – М.: ИСП РАН, 2013, 467 с.	https://www.ispras.ru/upload/uf/8b4/8b4258434b95638a21da0a1800181534.pdf

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

Электронно-библиотечная система Научно-технической библиотеки МИИТ http://library.miit.ru; Научно-электронная библиотека http://elibrary.ru.

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

Операционная система Linux.

Программные средства Docker, Docker Compose.

Docker образы HDFS, HADOOP, HIVE.

Среда программирования RStudio.

При организации обучения по дисциплине (модулю) с применением электронного обучения и дистанционных образовательных технологий необходим доступ каждого студента к информационным ресурсам – библиотечному фонду Университета, сетевым ресурсам и информационно-телекоммуникационной сети «Интернет».

В случае проведении занятий с применением электронного обучения и дистанционных образовательных технологий может понадобиться наличие следующего программного обеспечения (или их аналогов): ОС Windows, Microsoft Office, Интернет-браузер, Microsoft Teams и т.д.

В образовательном процессе, при проведении занятий с применением электронного обучения и дистанционных образовательных технологий, могут применяться следующие средства коммуникаций: ЭИОС РУТ(МИИТ), Microsoft Teams, электронная почта, скайп, Zoom, WhatsApp и т.п.

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

Учебные аудитории для проведения учебных занятий, оснащенные компьютерной техникой и наборами демонстрационного оборудования.

9. Форма промежуточной аттестации:

Зачет в 3 семестре.

10. Оценочные материалы.

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

Авторы:

доцент, доцент, к.н. кафедры
«Цифровые технологии управления
транспортными процессами» А.Ю. Павлов

Согласовано:

Заведующий кафедрой ЦТУТП В.Е. Нутович

Председатель учебно-методической
комиссии Н.А.Клычева