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.

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

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 больших данных, интегрируемых в корпоративные информационные системы,

методы хранения и обработки больших данных,

IT-сервисы и программные обеспечения, необходимые для работы с большими данных

Уметь:

находить данные для анализа в открытых источниках,

загружать данные, проводить предварительную чистку данных для анализа,

проводить анализ данных в зависимости от их типа

проводить установку Apache Hadoop

настраивать в минимальных требованиях Apache Hadoop

тестировать Apache Hadoop

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

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 №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 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. Проекты в области Больших Данных. Рассматриваемые вопросы: - CRISP-DM. Этапы. - Роли в проекте: Data Engineer, Data Analyst, Data Scientist.
2	Тема 2. Многомерный статистический анализ. Рассматриваемые вопросы: - Коэффициент корреляции. - Факторный анализ.
3	Тема 3. Методы многомерного статистического анализа. Рассматриваемые вопросы: - Оценка качества модели. - Многомерное шкалирование.
4	Тема 4. Методы многомерного статистического анализа. Рассматриваемые вопросы: - Дискриминантный анализ. - Методы контроля качества.
5	Тема 5. Методы многомерного статистического анализа. Рассматриваемые вопросы: - Иерархический кластерный анализ. - Метод K-means.
6	Тема 6. Технологии хранения и обработки больших данных. Рассматриваемые вопросы: - OLTP и OLAP системы. - DataWareHouse, DataLake и LakeHouse. - Модели Кимбалла и Инмона. - Схема Звезда и Снежинка. - Архитектура слоев хранения данных. - Методы загрузки данных: ETL и ELT.
7	Тема 7. Проектирование MPP-систем. Рассматриваемые вопросы: - DAMA DMBoK. - Data Vault и Anchor Modeling. - Вертикальное и горизонтальное масштабирование. - Вертикальное шардирование (партицирование). - Горизонтальное шардирование. - MPP – архитектура (Massively Parallel Processing). - Пакетная и потоковая обработка данных. - Лямбда – архитектура.
8	Тема 8. Хранение неструктурированных данных. Рассматриваемые вопросы: - ЦОД – TIER – RAID. - Проблема хранения неструктурированных данных: ACID и BASE. - Теорема CAP. - Реляционные СУБД и NoSQL.

№ п/п	Тематика лекционных занятий / краткое содержание
	- Модели данных NoSQL.
9	Тема 9. Парадигма MapReduce, Hadoop, HDFS Рассматриваемые вопросы: - Архитектура Hadoop. - HDFS. - Компоненты и планировщики YARN. - Концепция Map Reduce.
10	Тема 10. Apache Spark. Рассматриваемые вопросы: - Загрузка данных в Hadoop. - Spark Context. - Framework Apache Spark. - RDD. - Spark Streaming.
11	Тема 11. SQL поверх Hadoop. Рассматриваемые вопросы: - Архитектура Hive. - Развертывание MetaStore. - Trino.
12	Тема 12. Экосистема Hadoop. Рассматриваемые вопросы: - Apache Kafka. - Топики и партиции Kafka. - Apache ZooKeeper.
13	Тема 13. Технологии и методы анализа Big Data. Рассматриваемые вопросы: - Статистический анализ. - Метод смешения и интеграции данных. - Машинное обучение и Нейросети. - Data Mining. - Стратегия краудсорсинга. - Метод предиктивной аналитики. - Технология имитационного моделирования. - Визуализация аналитических данных – BI системы.
14	Тема 14. Языки программирования для Big Data. Рассматриваемые вопросы: - Архитектура Фон Неймана. - Закон Мура. - R, Python, SQL, Java, Scala, Julia, Go, C++.

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Методы многомерного статистического анализа. Рассматриваемые вопросы: - Практическое занятие по факторному анализу.

№ п/п	Тематика практических занятий/краткое содержание
2	Тема 2. Методы многомерного статистического анализа. Рассматриваемые вопросы: - Анализ данных с помощью Python. - Библиотеки Python: pandas, numpy, matplotlob, seaborn, scikit-learn
3	Тема 3. Методы многомерного статистического анализа. Рассматриваемые вопросы: - Анализ данных с помощью SQL.
4	Тема 4. Методы многомерного статистического анализа. Рассматриваемые вопросы: - Исследование датасета и кластеризация данных. - A/B анализ.
5	Тема 5. Технологии хранения и обработки больших данных. Рассматриваемые вопросы: - Работа с ETL сервисами.
6	Тема 6. Хранение неструктурированных данных. Рассматриваемые вопросы: - MongoDB. - Колочное хранение данных.
7	Тема 7. Парадигма MapReduce, Hadoop, HDFS. Рассматриваемые вопросы: - Word Count с использование Map Reduce. - Узлы кластера Hadoop.
8	Тема 8. Apache Spark. Рассматриваемые вопросы: - Анализ данных с помощью Apache Spark. - Работа с PySpark.
9	Тема 9. SQL поверх Hadoop. Рассматриваемые вопросы: - HiveQL. - Внутренние и внешние таблицы Hive.
10	Тема 10. Экосистема Hadoop Рассматриваемые вопросы: Компоненты кластера Apache Kafka. Коэффициент репликации Kafka.
11	Тема 11. Технологии и методы анализа Big Data Рассматриваемые вопросы: - Методы разметки данных. - Визуализация данных и работа с BI-инструментарием.
12	Тема 12. Языки программирования для Big Data. Рассматриваемые вопросы: - Сравнение производительности языков программирования.

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	Гудфеллоу, Я. Глубокое	https://e.lanbook.com/book/107901

	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	
2	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
3	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	https://e.lanbook.com/book/173018
4	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	https://e.lanbook.com/book/131710
5	Гультяева, Т. А. Методы статистического обучения в задачах регрессии и классификации : монография / Т. А. Гультяева, А. А. Попов, А. С. Саутин. — Новосибирск : НГТУ, 2016. — 323 с. — ISBN 978-5-7782-2817-7	https://e.lanbook.com/book/118291
6	Кук, Д. Машинное	https://e.lanbook.com/book/97353

	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	
7	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
8	Изучаем Spark: молниеносный анализ данных / Х. Карау, Э. Конвински, П. Венделл, М. Захария. — Москва : ДМК Пресс, 2015. — 304 с. — ISBN 978-5-97060-323-9	https://e.lanbook.com/book/90118?ysclid=lwkcbpunxi226114464

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).

Windows 7 и выше
Microsoft Office 2010
R consoler

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

Компьютер преподавателя
Компьютеры студентов
Монитор
Проектор Optoma W340UST
Экран для проектора
Маркерная доска

9. Форма промежуточной аттестации:

Зачет в 3 семестре.
Курсовая работа в 3 семестре.

10. Оценочные материалы.

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

Авторы:

директор	Б.В. Игольников
руководитель образовательной программы	О.Б. Проневич

Согласовано:

Директор	Б.В. Игольников
Руководитель образовательной программы	О.Б. Проневич
Председатель учебно-методической комиссии	Д.В. Паринов