МИНИСТЕРСТВО ТРАНСПОРТА РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«РОССИЙСКИЙ УНИВЕРСИТЕТ ТРАНСПОРТА»
(РУТ (МИИТ)

Рабочая программа дисциплины (модуля),
как компонент образовательной программы
базового высшего образования
по направлению подготовки
09.03.01 Информатика и вычислительная техника,
утвержденной первым проректором ФГАОУ ВО
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): Цифровая инженерия транспортных процессов

Форма обучения: Очная

Рабочая программа дисциплины (модуля) в виде электронного документа выгружена из единой корпоративной информационной системы управления университетом и соответствует оригиналу

Простая электронная подпись, выданная ФГАОУ ВО РУТ (МИИТ)
ID подписи: 937226
Подписал: руководитель образовательной программы
Проневич Ольга Борисовна
Дата: 07.09.2026

1. Общие сведения о дисциплине (модуле).

Целями освоения учебной дисциплины являются формирование у обучающихся профессиональных знаний и практических навыков в области анализа, интеграции, распределенного хранения, пакетной и потоковой обработки больших данных, а также подготовки эффективных визуализаций и аналитических отчетов для поддержки принятия решений.

Дисциплина предназначена для формирования и закрепления у обучающихся компетенций, необходимых для решения следующих профессиональных задач:

- проведение аналитических исследований с применением технологий больших данных;
- программная обработка структурированных, полуструктурированных и неструктурированных данных в существующей технологической инфраструктуре организации;
- проектирование воспроизводимых batch- и streaming-pipeline, обеспечение качества и производительности обработки данных;
- управление аналитическими работами, подготовка визуализаций, отчетов и рекомендаций по результатам исследования.

2. Планируемые результаты обучения по дисциплине (модулю).

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

ППК-ДА4 - Способен выявлять бизнес-проблемы или бизнес-возможности.

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

Знать:

принципы эффективной визуализации данных и требования к представлению результатов аналитического исследования;

инструменты создания аналитических отчетов по результатам обработки больших данных;

основы управления аналитическими работами, включая постановку задачи, планирование этапов, контроль качества и представление результатов;

основные свойства больших данных, жизненный цикл данных и архитектурные подходы к пакетной и потоковой обработке;

принципы распределенного хранения и вычислений, назначение хранилищ данных, озер данных и lakehouse-архитектур;

назначение структурированных, полуструктурированных и неструктурированных хранилищ, а также особенности SQL- и NoSQL-подходов;

методы обеспечения качества, согласованности, безопасности и воспроизводимости аналитических данных.

Уметь:

программировать на языках высокого уровня, ориентированных на работу с большими данными, для статистической обработки данных и работы с графикой;

обрабатывать разрозненные фрагменты данных в больших массивах, выполнять их очистку, преобразование, объединение, группировку и агрегирование;

работать с базами структурированных и неструктурированных данных, формировать аналитические запросы и извлекать необходимые выборки;

использовать Python, SQL и распределенные вычислительные фреймворки для пакетной обработки данных;

организовывать прием, хранение и обработку потоковых данных;

проводить разведочный и статистический анализ больших массивов данных с учетом ограничений вычислительных ресурсов;

выбирать способы визуализации и формировать аналитические отчеты, ориентированные на принятие управленческих решений;

оценивать качество и производительность аналитического процесса, выявлять узкие места и предлагать способы оптимизации;

планировать и документировать аналитические работы в соответствии с используемой в организации инфраструктурой.

Владеть:

навыками построения воспроизводимого аналитического pipeline для больших данных;

навыками применения Python, SQL, PySpark и инструментов визуализации при решении прикладных аналитических задач;

навыками работы с распределенными файловыми и объектными хранилищами, аналитическими СУБД и NoSQL-хранилищами;

навыками профилирования, контроля качества и подготовки данных к аналитическому исследованию;

навыками разработки аналитических отчетов и визуальных представлений результатов;

навыками управления этапами аналитического исследования и представления его результатов заинтересованным сторонам;

навыками применения технологий больших данных к транспортным, городским и организационным наборам данных.

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 №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 28 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. Введение в большие данные и задачи аналитических исследований Рассматриваемые вопросы: - понятие и ключевые характеристики больших данных; - источники больших данных в транспортных и городских системах; - типы аналитических задач и требования заинтересованных сторон; - жизненный цикл аналитического исследования.
2	Тема 2. Организация и управление аналитическими работами Рассматриваемые вопросы: - формулирование цели, гипотез и критериев результата; - планирование этапов и ресурсов аналитического проекта; - роли участников и управление аналитическим backlog; - контроль качества и документирование результатов.
3	Тема 3. Архитектуры хранения и обработки больших данных Рассматриваемые вопросы: - пакетная и потоковая обработка данных; - хранилище данных, озеро данных и lakehouse; - централизованные и распределенные архитектуры; - выбор архитектуры под прикладную задачу.
4	Тема 4. Распределенное хранение и форматы данных Рассматриваемые вопросы: - принципы распределенных файловых систем; - HDFS и объектные хранилища; - CSV, JSON, Parquet, ORC и Avro; - сжатие, схемы, метаданные и совместимость форматов.
5	Тема 5. Распределенные вычисления и модель MapReduce Рассматриваемые вопросы: - проблемы масштабирования вычислений; - модель MapReduce и ее основные этапы; - локальность данных и отказоустойчивость; - ограничения классических batch-подходов.
6	Тема 6. Apache Spark и распределенная обработка DataFrame Рассматриваемые вопросы: - архитектура Spark и компоненты кластера; - RDD, DataFrame и Dataset; - ленивые вычисления и план выполнения; - Spark SQL и применение PySpark.
7	Тема 7. Получение, интеграция и подготовка данных Рассматриваемые вопросы: - ETL и ELT-подходы; - интеграция файлов, API, баз данных и журналов событий; - очистка, нормализация и сопоставление сущностей; - идемпотентность и повторный запуск загрузок.
8	Тема 8. SQL- и NoSQL-хранилища в аналитических системах Рассматриваемые вопросы: - реляционные и аналитические СУБД; - документные, колоночные и key-value хранилища; - моделирование данных и выбор типа хранилища; - согласованность, доступность и масштабируемость.
9	Тема 9. Качество данных и управление данными Рассматриваемые вопросы:

№ п/п	Тематика лекционных занятий / краткое содержание
	- полнота, точность, согласованность и актуальность данных; - профилирование и правила валидации; - каталогизация, lineage и метаданные; - контроль доступа, безопасность и ответственность за данные.
10	Тема 10. Разведочный и статистический анализ больших данных Рассматриваемые вопросы: - описательные статистики и распределения; - выборки и агрегирование при больших объемах; - поиск взаимосвязей и проверка аналитических гипотез; - ограничения статистической интерпретации на больших данных.
11	Тема 11. Масштабируемая инженерия признаков и поиск аномалий Рассматриваемые вопросы: - обработка числовых, категориальных и временных признаков; - агрегаты, окна и временные интервалы; - методы выявления аномалий и контроля качества; - подготовка признаков для аналитики и машинного обучения.
12	Тема 12. Потоковая обработка и событийные данные Рассматриваемые вопросы: - событийная модель и брокеры сообщений; - Apache Kafka и организация потоков данных; - оконные вычисления и обработка опоздавших событий; - Spark Structured Streaming и сценарии real-time аналитики.
13	Тема 13. Эффективная визуализация больших данных Рассматриваемые вопросы: - выбор визуального представления под аналитическую задачу; - агрегирование и снижение визуальной перегрузки; - дашборды, аналитические отчеты и data storytelling; - ошибки визуализации и требования к интерпретируемости.
14	Тема 14. Облачные платформы и инфраструктура больших данных Рассматриваемые вопросы: - виртуальные машины, контейнеры и управляемые сервисы; - объектные хранилища и распределенные кластеры; - масштабирование ресурсов и оценка стоимости; - воспроизводимость и переносимость аналитической среды.
15	Тема 15. Производительность и оптимизация аналитических решений Рассматриваемые вопросы: - партиционирование, индексация и сортировка данных; - оптимизация соединений, агрегаций и запросов; - кэширование, управление памятью и вычислительными ресурсами; - профилирование и интерпретация планов выполнения.
16	Тема 16. Прикладные проекты больших данных в транспортной отрасли Рассматриваемые вопросы: - пассажиропотоки, телеметрия, ДТП и мобильность; - проектирование аналитической архитектуры транспортного кейса; - формирование аналитического отчета и рекомендаций; - оценка эффекта и презентация результатов исследования.

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Развертывание среды анализа больших данных Рассматриваемые вопросы: - организация проектной директории и виртуального окружения; - установка Python, JupyterLab, Docker и библиотек; - запуск контейнеров и проверка сетевого взаимодействия; - структура воспроизводимого аналитического проекта.
2	Тема 2. Постановка задачи и проектирование аналитического исследования Рассматриваемые вопросы: - перевод бизнес-потребности в аналитическую задачу; - определение источников, метрик и критериев результата; - составление плана аналитических работ; - оформление паспорта исследования.
3	Тема 3. Загрузка транспортных данных из файлов и API Рассматриваемые вопросы: - чтение CSV, JSON и архивов; - получение данных через REST API; - работа с кодировками и большими файлами; - контроль полноты и корректности загрузки.
4	Тема 4. Обработка больших файлов средствами Python Рассматриваемые вопросы: - чтение данных по частям; - потоковая обработка и генераторы; - снижение потребления памяти; - сохранение промежуточных результатов.
5	Тема 5. SQL-анализ данных в PostgreSQL Рассматриваемые вопросы: - создание таблиц и загрузка данных; - фильтрация, соединения и агрегирование; - оконные функции и аналитические запросы; - экспорт выборок в Python.
6	Тема 6. Аналитические запросы в ClickHouse Рассматриваемые вопросы: - особенности колоночного хранения; - создание таблиц семейства MergeTree; - агрегации и запросы по большим наборам данных; - сравнение производительности с реляционной СУБД.
7	Тема 7. Работа с документными и полуструктурированными данными Рассматриваемые вопросы: - моделирование документов JSON; - загрузка и выборка данных в MongoDB; - фильтрация, агрегации и вложенные структуры; - выбор SQL- или NoSQL-подхода.
8	Тема 8. Профилирование и контроль качества данных Рассматриваемые вопросы: - анализ типов, пропусков и уникальности; - поиск дубликатов и противоречий; - разработка правил валидации; - формирование отчета о качестве данных.
9	Тема 9. Форматы Parquet, ORC и Avro Рассматриваемые вопросы:

№ п/п	Тематика практических занятий/краткое содержание
	- преобразование данных между форматами; - сравнение размера и скорости чтения; - колоночное хранение и pushdown-фильтрация; - выбор формата под аналитическую задачу.
10	Тема 10. Основы PySpark и Spark DataFrame Рассматриваемые вопросы: - создание SparkSession; - загрузка и исследование DataFrame; - select, filter, withColumn и группировки; - ленивые вычисления и actions.
11	Тема 11. Распределенные преобразования данных в Spark Рассматриваемые вопросы: - очистка и стандартизация значений; - обработка пропусков и дубликатов; - пользовательские функции и встроенные функции Spark; - сохранение результатов в Parquet.
12	Тема 12. Соединения, агрегации и оконные функции в Spark Рассматриваемые вопросы: - join-операции и выбор стратегии соединения; - groupBy и агрегирующие функции; - оконные функции; - анализ плана выполнения.
13	Тема 13. Партиционирование и организация файлового слоя Рассматриваемые вопросы: - выбор ключей партиционирования; - проблема малых файлов; - repartition и coalesce; - оптимизация структуры каталога данных.
14	Тема 14. Работа с объектным хранилищем Рассматриваемые вопросы: - создание bucket и организация префиксов; - загрузка и получение объектов; - разграничение доступа; - использование объектного хранилища в аналитическом pipeline.
15	Тема 15. Разработка batch ETL-pipeline Рассматриваемые вопросы: - разделение этапов extract, transform и load; - параметризация и повторный запуск; - контроль ошибок и логирование; - сохранение витрины данных.
16	Тема 16. Основы Apache Kafka и событийной интеграции Рассматриваемые вопросы: - создание topic и настройка producer/consumer; - передача сообщений в формате JSON; - партиции, offsets и consumer groups; - контроль доставки и повторная обработка.
17	Тема 17. Потоковая обработка в Spark Structured Streaming Рассматриваемые вопросы: - подключение потокового источника; - оконные агрегации;

№ п/п	Тематика практических занятий/краткое содержание
	- watermark и опоздавшие события; - вывод результатов в приемник данных.
18	Тема 18. Разведочный анализ больших массивов данных Рассматриваемые вопросы: - агрегированные статистики и частотный анализ; - выборки и стратификация; - анализ распределений и взаимосвязей; - интерпретация результатов с учетом масштаба данных.
19	Тема 19. Статистические гипотезы на больших данных Рассматриваемые вопросы: - формулирование H0 и H1; - выбор критерия и уровня значимости; - оценка практической и статистической значимости; - представление результата в аналитическом отчете.
20	Тема 20. Выявление аномалий в транспортных данных Рассматриваемые вопросы: - формирование признаков и базовых правил; - статистические методы поиска выбросов; - Isolation Forest и кластерные подходы; - проверка и интерпретация найденных аномалий.
21	Тема 21. Анализ временных и географических данных Рассматриваемые вопросы: - подготовка временных меток и интервалов; - расчет динамики, сезонности и скользящих показателей; - агрегирование по территориям и объектам транспорта; - подготовка данных для карт и временных графиков.
22	Тема 22. Визуализация и формирование аналитического отчета Рассматриваемые вопросы: - выбор графиков и ключевых показателей; - визуализация агрегированных больших данных; - построение интерактивного отчета или дашборда; - формулирование выводов и рекомендаций.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение рекомендуемой литературы по архитектурам и технологиям больших данных.
2	Работа с официальной документацией Apache Spark, Hadoop, Kafka, ClickHouse, PostgreSQL и MongoDB.
3	Самостоятельная подготовка и профилирование наборов транспортных и городских данных.
4	Подготовка к текущему контролю и промежуточной аттестации.
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	Железнов, М. М. Методы и технологии обработки больших данных : учебно-методическое пособие / М. М. Железнов. — Москва : МИСИ – МГСУ, 2020. — 46 с. — ISBN 978-5-7264-2193-3.	https://e.lanbook.com/book/145102
2	Ланских, Ю. В. Введение в большие данные : учебное пособие / Ю. В. Ланских, В. Г. Ланских, К. В. Родионов. — Киров : ВятГУ, 2023. — 172 с.	https://e.lanbook.com/book/408566

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

http://library.miit.ru/ - электронно-библиотечная система научно-технической библиотеки МИИТ

https://superset.apache.org/user-docs/intro/#documentation – документация Apache Superset

https://www.postgresql.org/ - официальная страница PostgreSQL

https://yandex.cloud/ru/docs/datalens/ - официальная документация Yandex Datalens

https://docs.docker.com/ - официальная документация Docker

https://www.metabase.com/docs/latest/ - официальная документация Metabase

https://tableau.pro/ - официальный сайт Tableau с лучшими практиками проектирования дашбордов

https://revealthedata.com/blog/ - Блог о визуализации данных и BI-системах

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

MS Office (Excel, Word, PowerPoint)

Apache Superset

Yandex Datalens

Metabase

PostgreSQL

Docker
Apache Airflow
Браузер Chrome / Яндекс браузер
Draw.io

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

Компьютер преподавателя с доступом к сети Интернет
Компьютеры студентов с установленным Python-окружением и табличным процессором
Доступ к учебной базе событийных данных и экспериментальным выборкам
Доступ к PostgreSQL и BI-системе для визуализации результатов
Доступ к облачной виртуальной машине или Docker-контейнерам для воспроизводимого анализа
Проектор и 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:

руководитель образовательной программы О.Б. Проневич

Согласовано:

Директор Д.В. Паринов

Руководитель образовательной программы О.Б. Проневич

Председатель учебно-методической комиссии Д.В. Паринов