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)

Технологии хранения 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: 09.09.2026

1. Общие сведения о дисциплине (модуле).

Целями освоения учебной дисциплины являются формирование у обучающихся системных знаний и практических навыков в области технологий хранения и обработки больших данных для проведения аналитических исследований в транспортной отрасли. В рамках курса рассматриваются современные архитектуры хранения данных, распределенные файловые системы, NoSQL-базы данных, озера данных (Data Lake) и хранилища данных (Data Warehouse), методы статистического анализа данных, а также критерии выбора алгоритмов для различных задач анализа транспортных данных. Особое внимание уделяется работе с большими объемами транспортных телематических данных, логистических потоков и данных GPS/ГЛОНАСС, а также практическим аспектам применения технологий больших данных для решения прикладных задач отрасли.

Задачами освоения дисциплины являются:

Формирование теоретических знаний об архитектурах хранения больших данных, распределенных системах и моделях данных.

Освоение технологий и инструментов для работы с большими данными (Hadoop, Spark, Hive, HBase, Cassandra, MongoDB, ClickHouse).

Развитие практических навыков построения хранилищ данных и озер данных для транспортных задач.

Формирование компетенций по применению методов статистического анализа данных на больших объемах информации.

Освоение критериев выбора алгоритмов анализа данных в зависимости от типа решаемой транспортной задачи и характеристик данных.

Приобретение навыков проведения аналитических исследований с применением технологий больших данных в транспортной отрасли.

Дисциплина предназначена для формирования компетенций, необходимых для эффективного проведения аналитических исследований с применением технологий больших данных в транспортных организациях.

2. Планируемые результаты обучения по дисциплине (модулю).

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

ИД-1 - Способен формулировать и согласовывать требования к аналитическим исследованиям с применением технологий больших данных;

ППК-ДА6 - 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 статистического анализа данных, применяемые для обработки больших объемов транспортной информации.

Критерии выбора алгоритмов для различных задач анализа больших данных на транспорте.

Уметь:

Применять методы статистического анализа данных для исследования транспортных потоков, телематических данных и логистических процессов.

Выбирать оптимальные алгоритмы и технологии хранения данных в зависимости от типа решаемой задачи и характеристик транспортных данных.

Владеть:

Навыками работы с распределенными системами хранения и обработки данных (Hadoop, Spark) и NoSQL-базами данных.

Техниками построения хранилищ данных и озер данных для транспортных организаций

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 9 з.е. (324 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	№4
Контактная работа при проведении учебных занятий (всего):	144	64	80
В том числе:			
Занятия лекционного типа	48	16	32

Занятия семинарского типа	96	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 180 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 технологии хранения больших данных. Рассматриваемые вопросы: - Понятие больших данных (Big Data) и их характеристики (Volume, Velocity, Variety, Veracity, Value). - Особенности транспортных больших данных (GPS/ГЛОНАСС, логистика, телематика, биллинг). - Эволюция систем хранения данных: от файловых систем к распределенным. - Классификация систем управления базами данных. - Современные тренды и вызовы в хранении больших данных на транспорте.
2	Тема 2. Жизненный цикл данных в транспортных системах. Рассматриваемые вопросы: - Этапы жизненного цикла данных: сбор, хранение, обработка, анализ, архивация. - Особенности каждого этапа для транспортных данных. - Политики управления данными (Data Governance). - Роли и ответственность в управлении данными.
3	Тема 3. Архитектуры хранения больших данных: Data Warehouse. Рассматриваемые вопросы: - Концепция хранилища данных (Data Warehouse). - Схемы звезда и снежинка. - Преимущества и ограничения Data Warehouse. - Применение Data Warehouse для транспортной аналитики. - Примеры реализаций на транспорте.
4	Тема 4. Архитектуры хранения больших данных: Data Lake. Рассматриваемые вопросы: - Концепция озера данных (Data Lake). - Преимущества и ограничения Data Lake.

№ п/п	Тематика лекционных занятий / краткое содержание
	- Сравнение Data Lake и Data Warehouse. - Применение Data Lake для хранения сырых транспортных данных. - Кейсы использования на транспорте.
5	Тема 5. Архитектуры хранения больших данных: Data Lakehouse. Рассматриваемые вопросы: - Концепция Data Lakehouse: объединение Data Lake и Data Warehouse. - Преимущества и ограничения. - Технологии реализации Lakehouse (Delta Lake, Iceberg, Hudi). - Применение для транспортных задач. - Перспективы развития Lakehouse на транспорте.
6	Тема 6. Сравнительный анализ архитектур хранения для транспортных задач. Рассматриваемые вопросы: - Критерии выбора архитектуры хранения. - Сравнительный анализ Data Warehouse, Data Lake, Lakehouse. - Выбор архитектуры в зависимости от типа транспортной задачи. - Гибридные подходы и их применение. - Рекомендации для транспортных организаций.
7	Тема 7. Распределенные файловые системы: HDFS. Рассматриваемые вопросы: - Понятие распределенной файловой системы. - HDFS: архитектура и принципы работы (NameNode, DataNode). - Репликация и отказоустойчивость в HDFS. - Работа с большими файлами в HDFS. - Интеграция HDFS с инструментами обработки.
8	Тема 8. Экосистема Hadoop. Рассматриваемые вопросы: - Обзор экосистемы Hadoop. - YARN: управление ресурсами и планирование задач. - Hive: SQL-интерфейс для Hadoop. - HBase: NoSQL-база данных на Hadoop. - Применение экосистемы Hadoop для транспортных данных.

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Установка и настройка среды для работы с большими данными. Рассматриваемые вопросы: - Установка Docker и Docker Compose. - Настройка окружения для Hadoop, Spark, ClickHouse. - Создание кластера в контейнерах. - Проверка работоспособности всех компонентов. - Первые шаги в командной строке.
2	Тема 2. Введение в экосистему Hadoop. Рассматриваемые вопросы: - Знакомство с архитектурой HDFS. - Базовые команды работы с HDFS (ls, put, get, cat, mkdir). - Загрузка и выгрузка данных в HDFS.

№ п/п	Тематика практических занятий/краткое содержание
	- Работа с правами доступа в HDFS. - Мониторинг HDFS.
3	Тема 3. Продвинутая работа с HDFS. Рассматриваемые вопросы: - Репликация файлов и управление репликацией. - Работа с большими файлами (блоки, партиции). - Сжатие данных в HDFS (gzip, snappy, lzo). - Настройка квот и пространства имен. - Мониторинг состояния кластера.
4	Тема 4. Введение в Apache Hive. Рассматриваемые вопросы: - Установка и настройка Hive. - Создание таблиц в Hive (внутренние и внешние). - Загрузка данных в Hive из HDFS. - Выполнение SQL-запросов в Hive (SELECT, GROUP BY, JOIN). - Практика: загрузка транспортного датасета.
5	Тема 5. Продвинутая работа с Apache Hive. Рассматриваемые вопросы: - Секционирование таблиц в Hive. - Использование ведра (bucketing) для оптимизации. - Сложные запросы с подзапросами и CTE. - Интеграция Hive с другими инструментами. - Оптимизация производительности запросов.
6	Тема 6. Введение в Apache HBase. Рассматриваемые вопросы: - Архитектура HBase (Region Server, HMaster, ZooKeeper). - Создание таблиц в HBase. - Вставка и извлечение данных (put, get, scan). - Работа с фильтрами и сканированием. - Практика: хранение телематических данных.
7	Тема 7. Интеграция HBase с Hive. Рассматриваемые вопросы: - Создание внешних таблиц в Hive над HBase. - Выполнение SQL-запросов к данным в HBase. - Сравнение производительности Hive и HBase. - Сценарии использования HBase на транспорте. - Практика: гибридный запрос.
8	Тема 8. Введение в MongoDB: установка и базовые операции. Рассматриваемые вопросы: - Установка и настройка MongoDB. - Структура документов в MongoDB. - CRUD-операции в MongoDB (insert, find, update, delete). - Создание индексов для оптимизации запросов. - Практика: загрузка транспортных данных.
9	Тема 9. Агрегационные запросы в MongoDB. Рассматриваемые вопросы: - Конвейеры агрегации (aggregation pipeline). - Операторы агрегации: match,match,group, project,project,sort, $limit. - Сложные агрегации с несколькими этапами.

№ п/п	Тематика практических занятий/краткое содержание
	- Анализ транспортных данных с агрегациями. - Практика: анализ пассажиропотоков.
10	Тема 10. Оптимизация MongoDB для транспортных данных. Рассматриваемые вопросы: - Настройка производительности MongoDB. - Использование составных индексов. - Оптимизация запросов с explain(). - Шардирование и репликация в MongoDB. - Практика: оптимизация запросов.
11	Тема 11. Введение в ClickHouse: установка и базовые операции. Рассматриваемые вопросы: - Установка и настройка ClickHouse. - Создание таблиц с различными движками (MergeTree, Log). - Загрузка данных в ClickHouse. - Выполнение базовых запросов (SELECT, WHERE, ORDER BY). - Практика: загрузка транспортного датасета.
12	Тема 12. Продвинутая работа с ClickHouse. Рассматриваемые вопросы: - Создание таблиц с партиционированием и сортировкой. - Использование материализованных представлений. - Работа с временными рядами в ClickHouse. - Оптимизация запросов с EXPLAIN. - Практика: сложные аналитические запросы.
13	Тема 13. Аналитика транспортных данных в ClickHouse. Рассматриваемые вопросы: - Агрегация больших объемов данных. - Группировка по времени и категориям. - Вычисление скользящих средних и трендов. - Анализ пассажиропотоков и загрузки. - Визуализация результатов аналитики.
14	Тема 14. Сравнительный анализ NoSQL-решений для транспортных данных. Рассматриваемые вопросы: - Сравнение MongoDB, ClickHouse, Cassandra. - Критерии выбора для различных задач. - Тестирование производительности на транспортных данных. - Выбор решения для конкретной транспортной задачи. - Документирование результатов сравнения.
15	Тема 15. Введение в Apache Cassandra. Рассматриваемые вопросы: - Архитектура Cassandra: принципы распределения, репликации, согласованности. - Установка и настройка Cassandra. - Создание keyspace и таблиц. - CRUD-операции в Cassandra. - Практика: работа с транспортными данными.
16	Тема 16. Продвинутая работа с Cassandra. Рассматриваемые вопросы: - Моделирование данных для Cassandra (партиционирование, кластеризация). - Оптимизация запросов в Cassandra. - Настройка согласованности и доступности.

№ п/п	Тематика практических занятий/краткое содержание
	- Масштабирование Cassandra-кластера. - Практика: оптимизация запросов на транспортных данных.
17	Тема 17. Введение в Apache Spark. Рассматриваемые вопросы: - Установка и настройка Spark. - Архитектура Spark: драйвер, исполнители. - Spark Session и Spark Context. - Основные операции с DataFrames (select, filter, groupBy). - Практика: загрузка и первичный анализ данных.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение литературы и документации по Hadoop, Spark, Hive, HBase.
2	Работа с официальной документацией NoSQL-баз данных (MongoDB, Cassandra, ClickHouse).
3	Изучение методов статистического анализа больших данных.
4	Выполнение домашних заданий по работе с распределенными системами хранения
5	Подготовка презентации по результатам проекта для защиты.
6	Подготовка к промежуточной аттестации.
7	Выполнение курсовой работы.
8	Подготовка к промежуточной аттестации.
9	Подготовка к текущему контролю.

4.4. Примерный перечень тем курсовых работ

Проектирование и реализация озера данных (Data Lake) для хранения GPS/ГЛОНАСС-данных транспортного предприятия.

Сравнительный анализ производительности ClickHouse и MongoDB для аналитики пассажиропотоков.

Разработка ETL-пайплайна для загрузки и трансформации транспортных телематических данных с использованием Apache Airflow и Spark.

Кластерный анализ маршрутов общественного транспорта на основе данных GPS с использованием Spark MLlib.

Построение хранилища данных (Data Warehouse) для агрегирования и анализа логистических показателей транспортной компании.

Разработка системы мониторинга качества данных в озере данных транспортной организации с использованием Great Expectations.

Прогнозирование пассажиропотоков городского транспорта с использованием Spark ML и статистических методов.

Обнаружение аномалий в данных телеметрии подвижного состава с использованием Isolation Forest в Spark.

Проектирование архитектуры Lakehouse для хранения и анализа мультимодальных транспортных данных с использованием Delta Lake.

Разработка интерактивного дашборда для мониторинга транспортных потоков на основе данных из ClickHouse и Apache Superset.

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

№ п/п	Библиографическое описание	Место доступа
1	Исабекова, О. А. Технологии формирования межотраслевого и межгосударственного единого информационного пространства : учебно-методическое пособие / О. А. Исабекова, Е. М. Дворникова. — Москва : РТУ МИРЭА, 2024. — 68 с. — ISBN 978-5-7339-2375-8.	https://e.lanbook.com/book/464624
2	Системы искусственного интеллекта и большие данные. Раздел «Большие данные» : учебное пособие / А. А. Парамонов, И. А. Юрченков, Б. А. Крынецкий, И. В. Есипов. — Москва : РТУ МИРЭА, 2025. — 147 с. — ISBN 978-5-7339-2499-1.	https://e.lanbook.com/book/493550

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 3 семестре.
Курсовая работа в 4 семестре.
Экзамен в 4 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