

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)

Инженерия транспортных данных: ETL/ELT-процессы

Направление подготовки: 09.03.01 Информатика и вычислительная техника

Направленность (профиль): IT-сервисы и технологии обработки данных на транспорте

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: директор центра Проневич Ольга Борисовна

Дата: 10.09.2026

1. Общие сведения о дисциплине (модуле).

Цель дисциплины: Формирование компетенций по построению сквозных конвейеров обработки транспортных данных (ETL/ELT) — от извлечения «сырой» телематики и расписаний до загрузки в аналитические хранилища для управления мобильностью.

Задачи дисциплины: изучение источников и протоколов транспортных данных (GPS/ГЛОНАСС, CAN, GTFS, АСКП); освоение методов потоковой и пакетной очистки, валидации и геопривязки треков; приобретение навыков проектирования моделей данных (DWH, Data Lake); освоение инструментов оркестрации (Airflow) и методов обеспечения качества данных в транспортных экосистемах.

2. Планируемые результаты обучения по дисциплине (модулю).

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

АД-2 - Способен проектировать, реализовывать и оптимизировать витрины данных для поддержки бизнес-аналитики;

ППК-Р2 - Способен проверять работоспособность и проводить рефакторинг кода программного обеспечения;

ППК-Р6 - Способен участвовать в промышленной разработке программного обеспечения.

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

Знать:

- Знает архитектурные паттерны и инструменты ETL/ELT для построения конвейеров обработки транспортных данных.
- Знает методы обработки инкрементальных данных (CDC, логгирование изменений) для организации потоковой и пакетной загрузки.

Уметь:

- Умеет разрабатывать скрипты извлечения, трансформации и загрузки данных с использованием Python и библиотек Pandas/NumPy.
- Умеет настраивать автоматическое обновление витрин данных по расписанию или событиям с использованием инструментов оркестрации (Apache Airflow).

Владеть:

- Владеет навыками проектирования и реализации ETL/ELT-пайплайнов для транспортных данных с применением методов инкрементальной загрузки (CDC) и автоматизацией обновления витрин.

- Владеет навыками инспекции программного кода ETL-скриптов для выявления ошибок, критических мест и проблем производительности, не обнаруженных на ранних стадиях разработки.

3. Объем дисциплины (модуля).

3.1. Общая трудоемкость дисциплины (модуля).

Общая трудоемкость дисциплины (модуля) составляет 5 з.е. (180 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 №4
Контактная работа при проведении учебных занятий (всего):	96	96
В том числе:		
Занятия лекционного типа	32	32
Занятия семинарского типа	64	64

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 84 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	<p>Тема 1. Введение в инженериию транспортных данных</p> <p>Рассматриваются вопросы:</p> <p>Роль данных в цифровой трансформации транспорта (Mobility as a Data). Источники данных: GPS/ГЛОНАСС-трекеры, CAN-шины, системы оплаты проезда (АСКП), дорожные контроллеры. Жизненный цикл данных. Отличие ETL от ELT в контексте транспортных потоков.</p>
2	<p>Тема 2. Архитектура транспортных дата-пайплайнов (часть 1)</p> <p>Рассматриваются вопросы:</p> <p>Пакетная (Batch) обработка данных расписаний и графика движения. Поточковая (Streaming) обработка телематики в реальном времени. Lambda и Карра-архитектуры. Принципы отказоустойчивости при потере сигнала транспортного средства.</p>
3	<p>Тема 3. Источники транспортных данных и протоколы сбора</p> <p>Рассматриваются вопросы:</p> <p>Стандарты передачи данных: NMEA (GPS-треки), GTFS (расписания общественного транспорта), DATEX II (дорожные события). Бортовые системы (FMS в авиации, АСУД в метро), IoT-сенсоры на дорожной инфраструктуре.</p>
4	<p>Тема 4. Извлечение данных (Extract) из неоднородных транспортных систем</p> <p>Рассматриваются вопросы:</p> <p>Полный и инкрементальный экстракт. Технологии Change Data Capture (CDC) для биллинга в платных дорогах. Работа с сырыми логами серверов бронирования. Скрипты-коннекторы к API перевозчиков (РЖД, авиакомпаний).</p>
5	<p>Тема 5. Трансформация геопространственных транспортных данных</p> <p>Рассматриваются вопросы:</p> <p>Обработка треков: сглаживание шума, привязка к дорожному графу (Map Matching). Преобразование координат (WGS-84 / МСК). Расчет производных величин (скорость, ускорение, пробег). Геозонирование (прибытие на остановку, выезд из депо).</p>
6	<p>Тема 6. Трансформация расписаний и маршрутных сетей (GTFS-спецификация)</p> <p>Рассматриваются вопросы:</p> <p>Парсинг и валидация GTFS-фидов (расписания). Создание справочников остановок, маршрутов, перегонов. Построение графа мультимодальных пересадок. Контроль качества: поиск "разрывов" в плечах маршрутов.</p>
7	<p>Тема 7. Очистка и нормализация данных «сырой телематики»</p> <p>Рассматриваются вопросы:</p> <p>Детекция "битых" треков (дрейф координат, нулевые пробеги). Фильтрация аномалий: скорость > физического предела вида транспорта. Механизмы дедупликации данных с валидаторов. Таксономия ошибок навигационного оборудования.</p>
8	<p>Тема 8. ETL для транспортной аналитики и BI (Business Intelligence)</p> <p>Рассматриваются вопросы:</p> <p>Расчет транспортной работы: пассажиро-километры, рейсообороты. Агрегация данных по видам тяги. Модели данных «Звезда» и «Снежинка» для аналитики депо. Интеграция данных топливных карт с путевыми листами.</p>
9	<p>Тема 9. ELT-подход с использованием Data Lake для транспортного сектора</p> <p>Рассматриваются вопросы:</p> <p>Хранение неструктурированных данных (видеопотоки распознавания номеров, данные LiDAR). Архитектура Data Vault для гибкой истории изменений транспортной инфраструктуры. Баланс между Data Lake и DWH при анализе пассажиропотока.</p>
10	<p>Тема 10. Управление мастер-данными (MDM) транспортной инфраструктуры</p> <p>Рассматриваются вопросы:</p> <p>Ведение "Золотой записи" остановочного пункта. Связывание разных идентификаторов (Osm_id,</p>

№ п/п	Тематика лекционных занятий / краткое содержание
	справочник региона, внутренний ID перевозчика). Медленно меняющиеся измерения (SCD) для тарифных справочников.
11	Тема 11. Оркестрация сложных транспортных пайплайнов Рассматриваются вопросы: Построение DAG-ов (направленных ациклических графов) для ночной загрузки отчетов. Обработка зависимостей: сначала геокодер, потом матчинг. Мониторинг SLA загрузки данных о ремонтах пути. Инструменты: Apache Airflow, Dagster.
12	Тема 12. Поточковая обработка и Complex Event Processing (CEP) на транспорте Рассматриваются вопросы: Архитектура брокеров сообщений (Kafka). Анализ потоков в реальном времени: определение затора/пробки до окончания поездки. Событийная модель ДТП: автоматическая фиксация резкого торможения. Окна событий.

4.2. Занятия семинарского типа.

Практические занятия

№ п/п	Тематика практических занятий/краткое содержание
1	Тема 1. Проектирование источника: симуляция движения транспортного средства Рассматриваются вопросы: Написание скрипта-симулятора GPS-трека автобуса по маршруту. Генерация NMEA-строк с типичными ошибками (потеря сигнала, скачки координат). Настройка логирования в файл.
2	Тема 2. Развертывание среды для ETL (Docker, Jupyter, Airflow) Рассматриваются вопросы: Поднятие инфраструктурного стека в Docker Compose. Установка Python-библиотек (Pandas, GeoPandas, SQLAlchemy). Инициализация базы данных PostgreSQL/PostGIS для хранения рейсов.
3	Тема 3. Extract: Подключение к REST API транспортного агрегатора Рассматриваются вопросы: Разбор документации API открытых данных (на примере данных МосТранса или Mapbox). Авторизация по API-ключу. Скрипт выгрузки прогноза погоды и дорожной ситуации для заданного плеча.
4	Тема 4. Extract: Парсинг статичных данных GTFS-расписания (часть 1) Рассматриваются вопросы: Чтение файлов agency.txt, stops.txt, routes.txt. Импорт в реляционные таблицы. Создание первичных ключей и связей. Практика работы с ZIP-архивом.
5	Тема 5. Extract: Парсинг GTFS-расписания (часть 2) и построение маршрутной сети Рассматриваются вопросы: Обработка trips.txt и stop_times.txt. Восстановление последовательности остановок. Визуализация маршрута на карте средствами Keplergl/Folium.
6	Тема 6. Extract: Считывание потока с бортового диагностического разъема (CAN) Рассматриваются вопросы: Подключение к файлу лога CAN-шины (формат .csv/asc). Фильтрация PID-кодов (обороты двигателя, расход топлива). Привязка показаний к временной шкале.
7	Тема 7. Transform: Очистка «сырых» треков (с использованием Pandas) Рассматриваются вопросы: Удаление выбросов координат (метод межквартильного размаха). Интерполяция пропусков в треке (линейная/сплайновая). Определение и вырезка стоянок.

№ п/п	Тематика практических занятий/краткое содержание
8	Тема 8. Transform: Map Matching – привязка трека к дорожному графу Рассматриваются вопросы: Использование библиотеки OSRM/Valhalla (локально). Алгоритм проекции "битых" GPS-точек на ребро дороги. Анализ процента успешной привязки к графу OSM.
9	Тема 9. Transform: Расчет транспортной работы рейса Рассматриваются вопросы: SQL-запросы на соединение данных трека и справочника маршрутов. Расчет линейного пробега и времени в движении. Выявление отклонений от планового расписания.
10	Тема 10. Transform: Нормализация адресных данных и геокодинг Рассматриваются вопросы: Приведение адресов депо к единому стандарту (КЛАДР/ФИАС). Прямое и обратное геокодирование через Nominatim. Создание таблицы почтовых адресов транспортных узлов
11	Тема 11. Load: Загрузка данных в хранилище (DWH) по схеме «Звезда» Рассматриваются вопросы: Проектирование таблицы фактов «Поездки» и измерений «Время», «Водитель», «Транспортное средство». Написание ETL-процедуры заполнения измерения медленно меняющихся данных (SCD Type 2).
12	Тема 12. Load: Инкрементальная загрузка истории транзакций (CDC) Тема 12. Load: Инкрементальная загрузка истории транзакций (CDC) Рассматриваются вопросы: Эмуляция журнала транзакций валидатора (оплата проезда). Реализация захвата изменений через хеш-суммы строк. Загрузка дельты слиянием (Merge/Upsert). Рассматриваются вопросы: Эмуляция журнала транзакций валидатора (оплата проезда). Реализация захвата изменений через хеш-суммы строк. Загрузка дельты слиянием (Merge/Upsert).
13	Тема 13. Визуализация качества данных (Data Quality Dashboard) Рассматриваются вопросы: Подсчет метрик: количество нулевых координат, дубликатов записей транзакций. Визуализация в Superset/Grafana. Настройка алертов на резкое падение количества треков.
14	Тема 14. Оркестрация: Создание первого DAG в Apache Airflow Рассматриваются вопросы: Описание графа задач: Extract -> Clean -> Load. Передача данных через XCom. Планирование запуска каждые 30 минут для эмуляции обновления расписания.
15	Тема 15. Ветвление и обработка ошибок в Airflow DAG Рассматриваются вопросы: BranchOperator: проверка наличия исходного файла. Механизм повторных попыток (retries) при сетевом сбое API перевозчика. Отправка уведомлений о падении пайплайна.
16	Тема 16. Поточковая обработка: Основы Apache Kafka Рассматриваются вопросы: Развертывание Kafka-брокера. Создание топика «gps_coordinates». Написание Producer на Python, симулирующего отправку координат с частотой 1 Гц.
17	Тема 17. Поточковая обработка: Создание Consumer и оконные функции Рассматриваются вопросы: Чтение потока с помощью Faust. Расчет скользящей средней скорости за 5 минут. Обнаружение остановки ТС по падению скорости (сессионные окна).
18	Тема 18. ELT: Трансформация внутри хранилища с помощью dbt (data build tool) Рассматриваются вопросы: Создание моделей dbt для сырых слоев. Материализация представлений (view) для витрины

№ п/п	Тематика практических занятий/краткое содержание
	опозданий рейсов. Тестирование связей ссылочной целостности «нет рейса без маршрута» (dbt tests).
19	Тема 19. ELT: Применение SQL-аналитики для поиска безбилетников Рассматриваются вопросы: Соединение данных валидации входа и выхода (реконструкция поездки). Логика выявления дисбаланса (вошел, но не оплатил в течение N минут). Расчет коэффициента потерь на остановке.
20	Тема 20. Работа с геоданными в PostGIS Рассматриваются вопросы: Пространственные индексы. SQL-запрос: найти все ТС, попавшие в геозону «Депо». Функции ST_Distance, ST_Contains. Расчет ближайшей остановки к точке инцидента.
21	Тема 21. Анализ пассажиропотока (Матрица корреспонденций) Рассматриваются вопросы: Построение таблицы входа-выхода на основе данных АСКП. Агрегация пассажиропотока между остановками. Экспорт матрицы в формат для транспортного моделирования (Visum)
22	Тема 22. Анализ цепочек поездок (Trip Chaining) Рассматриваются вопросы: Склеивание записей банковских карт/социальных карт в рамках одной корреспонденции. Классификация на пересадки и конечную цель. Расчет времени пересадки на ТПУ.
23	Тема 23. Feature Engineering: Подготовка данных для прогноза ETA Рассматриваются вопросы: Генерация признаков (время суток, день недели, загруженность дороги). Учет исторического времени проезда перегона. Создание плоской таблицы для обучения модели.
24	Тема 24. Feature Store: Регистрация и хранение транспортных признаков Рассматриваются вопросы: Сохранение рассчитанных признаков пробок (уровень сервиса дороги LOS) в Feature Store (Feast). Обеспечение консистентности тренировочных и инференс-данных.
25	Тема 25. Мониторинг дрейфа данных (Data Drift) Рассматриваются вопросы: Сравнение распределений скорости вчера и сегодня (тест Колмогорова-Смирнова). Настройка Evidently AI для отчета о дрейфе. Анализ влияния изменения расписания на модель.

4.3. Самостоятельная работа обучающихся.

№ п/п	Вид самостоятельной работы
1	Изучение литературы по ETL/ELT и Data Engineering.
2	Изучение документации по Apache Airflow.
3	Разработка ETL-пайплайна для выбранного транспортного сценария.
4	Написание Unit-тестов для своих скриптов трансформации.
5	Подготовка к текущему контролю и защите проекта.
6	Подготовка к промежуточной аттестации.
7	Подготовка к текущему контролю.

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

№ п/п	Библиографическое описание	Место доступа
1	Митина, О. А. Технологии организации, обработки и хранения статистических данных : учебное пособие / О. А. Митина, И. А. Юрченков. — Москва : РТУ МИРЭА, 2019. — 163 с.	https://e.lanbook.com/book/171511
2	Кудрявцева, И. Г. Основы бизнес-аналитики : учебно-методическое пособие / И. Г. Кудрявцева. — Москва : РТУ МИРЭА, 2025. — 237 с. — ISBN 978-5-7339-2548-6.	https://e.lanbook.com/book/498062

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

MS Office (Excel, Word)

Notepad++

Браузер Chrome

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

Инструменты визуального моделирования

Офисный пакет приложений

Система контроля версий Git

Браузер

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:

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

Директор ЦС IT - программ ВИШ

О.Б. Проневич

Председатель учебно-методической
комиссии

Д.В. Паринов