Департамент образования и науки города Москвы

Государственное автономное образовательное учреждение

высшего образования города Москвы

«Московский городской педагогический университет»

Институт цифрового образования

Департамент информатики управления и технологий

Мошенина Елена Дмитриевна БД-241м

**Лабораторная работа 4.1. Установка и настройка ETL-инструмента.**

**Вариант 14**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Руководитель дисциплины:

Босенко Т.М., доцент департамента

информатики, управления и технологий,

кандидат технических наук

Москва

2025

**Цель работы**: изучение основных принципов работы с ETL-инструментами на примере Pentaho [Data](http://95.131.149.21/moodle/mod/folder/view.php?id=1500) Integration (PDI), настройка конвейера обработки данных, фильтрация и замена значений в Excel-файле, а также выгрузка обработанных данных в базу данных MySQL/PostgreSQL.

**Основная часть**

**Лабораторная работа 4.1. Установка и настройка ETL-инструмента.**

 Настроить среду для работы с Pentaho Data Integration (PDI):

Запуск виртуальной машины с Ubuntu 22.04 в VirtualBox.

Проверка установки Java и WebKitGTK.

Развертывание Pentaho Data Integration.

Запуск Pentaho Data Integration

Переход в каталог data-integration

cd ~/Downloads/data-integration/

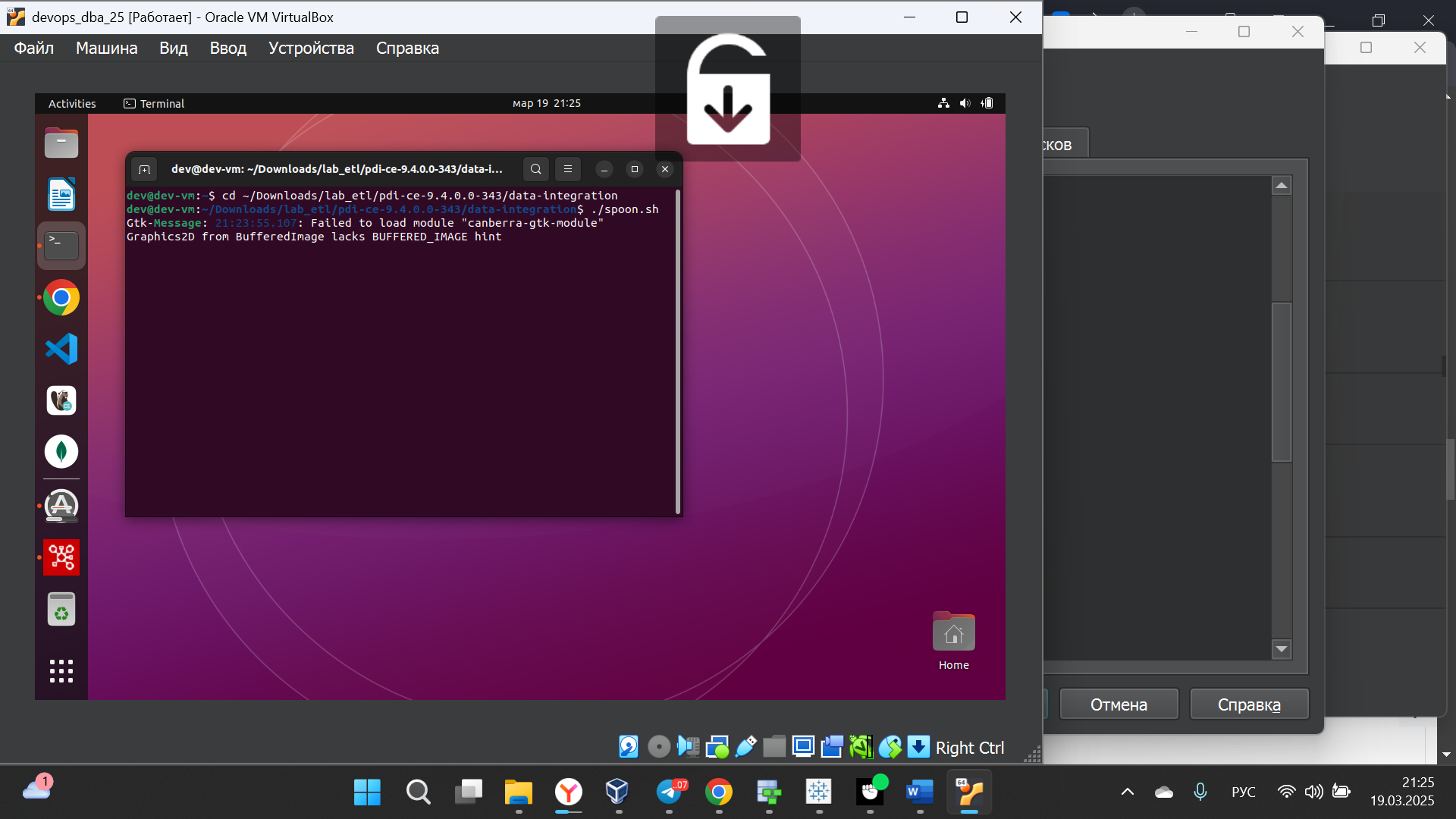
Делаем исполняемым файл Spoon (если нужно)

chmod +x spoon.sh

Запускаем Pentaho Spoon

./spoon.sh

Запускаем программу с помощью следующих команд:



Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Открываем нужный источник:

Изображение выглядит как текст, программное обеспечение, Значок на компьютере, Мультимедийное программное обеспечение

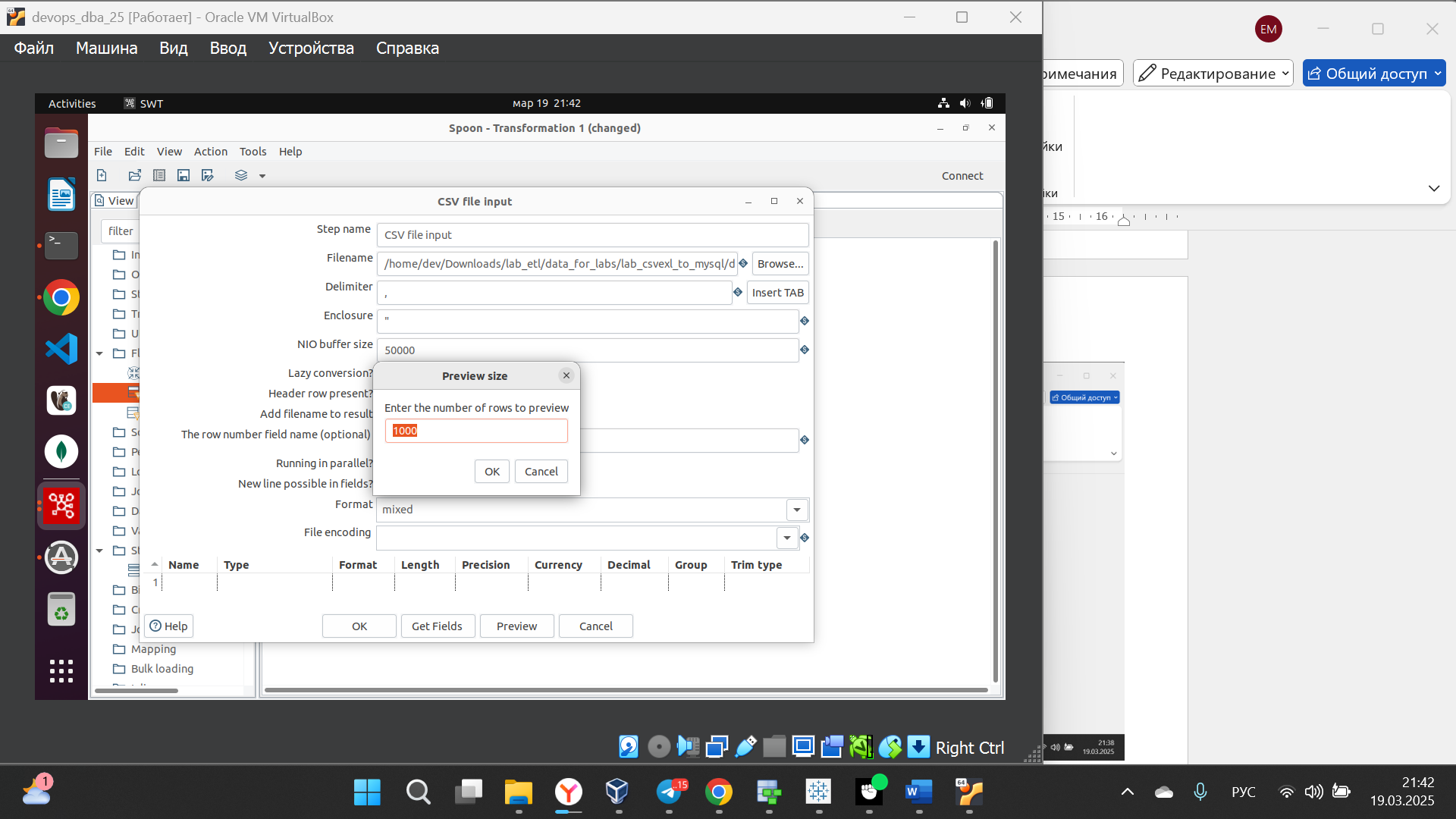
Контент, сгенерированный ИИ, может содержать ошибки.

Создаем трансформацию:

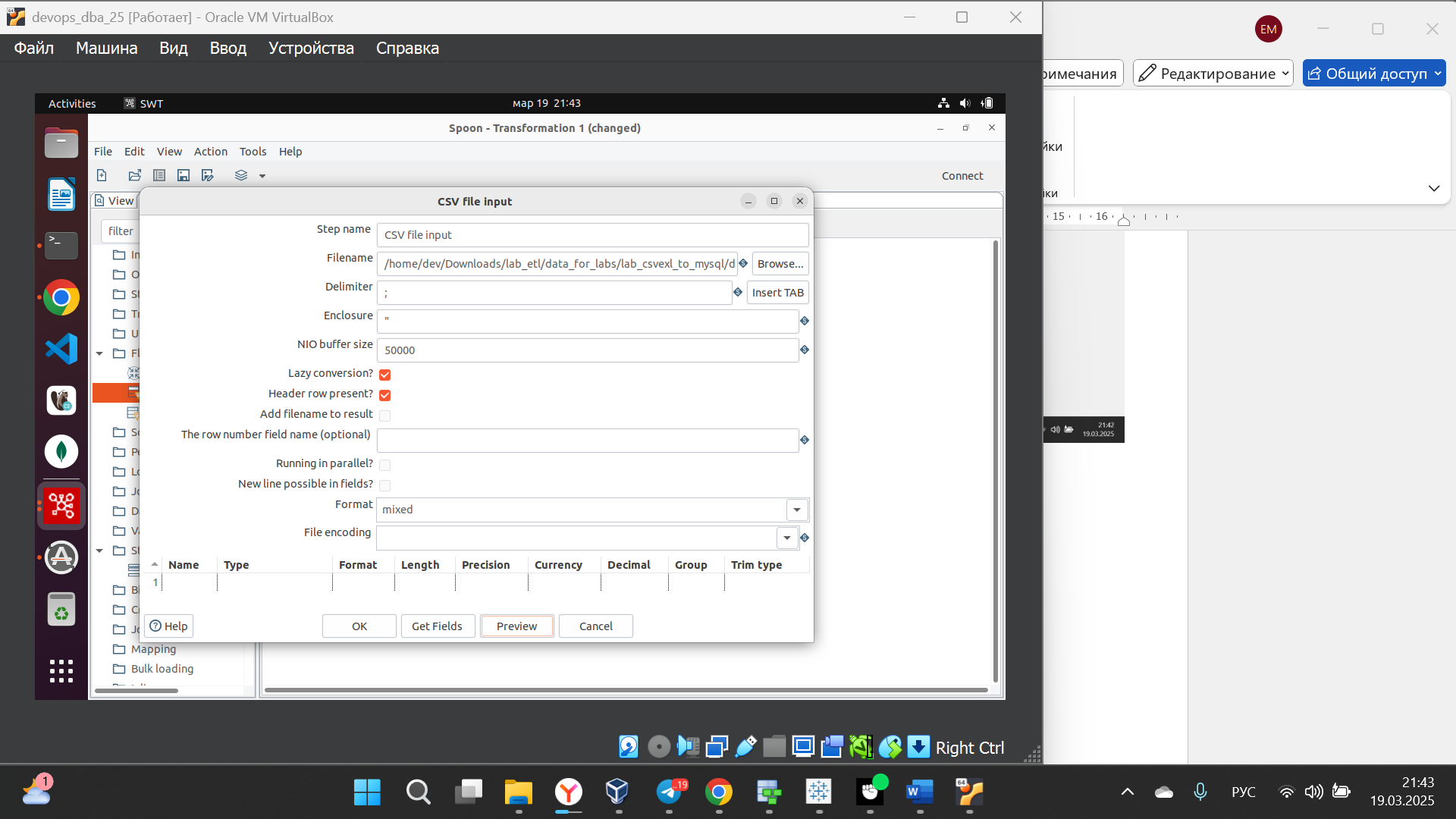
Изображение выглядит как текст, программное обеспечение, Мультимедийное программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Проверим, отработает ли первые 1000 значений:



Получилось плохо, возможно у нас другой разделитель, поменяем и проверим:



Теперь смотрим output. Выбираем, к какой базе данных буем отправлять данные.

Откроем такую схему:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Нужно будет создать новое соединение, но перед этим изучим старое:

Изображение выглядит как текст, программное обеспечение, Значок на компьютере, Мультимедийное программное обеспечение

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Заменить поля на необходимые, преобразуем и делаем с маленькой буквы и нижним подчеркиванием, меняем типы данных, тогда будет читаемо для СУБД.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, программное обеспечение, снимок экрана, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Проверяем корректность соединения. Все работает:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

Итог:

Изображение выглядит как текст, снимок экрана, программное обеспечение, Значок на компьютере

Контент, сгенерированный ИИ, может содержать ошибки.

**Вывод.** В ходе работы были изучены основные принципы работы с ETL-инструментами на примере Pentaho Data Integration (PDI). Были освоены процессы настройки конвейера обработки данных, включая фильтрацию и замену значений в Excel-файле. Кроме того, была реализована выгрузка обработанных данных в базу данных MySQL или PostgreSQL.

Это позволило получить практические навыки работы с ETL-процессами, понять, как эффективно обрабатывать и преобразовывать данные из различных источников, а также научиться интегрировать их в целевые системы для дальнейшего анализа и использования. Полученные знания и навыки могут быть полезны при работе с большими объёмами данных и в задачах, требующих автоматизации процессов извлечения, преобразования и загрузки данных.