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

Кузьмина Дарья Юрьевна БД-241м

**Практическая работа**

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

**Вариант 11**

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,

доктор экономических наук

Москва

2025

Содержание

[Введение 2](#_Toc193677793)

[Основная часть 4](#_Toc193677794)

[Заключение 9](#_Toc193677795)

## Введение

**Создание конвейеров данных**

**Цель работы**: изучение основных принципов работы с ETL-инструментами на примере Pentaho [Data](http://95.31.0.249/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.

**Условие выполнения работы**:

Работа выполняется в образе **Ubuntu 22.04 (.ova) для VirtualBox 7.0**<https://disk.yandex.ru/d/gagWU_zn1erR8g>, в котором предварительно установлены все необходимые компоненты для работы с Pentaho [Data](http://95.31.0.249/moodle/mod/folder/view.php?id=1500) Integration, либо проводится установка окружения в ОС Linux.

**Задачи**

-        **Настроить среду для работы с Pentaho**[**Data**](http://95.31.0.249/moodle/mod/folder/view.php?id=1500)**Integration (PDI)**:

Запуск виртуальной машины с **Ubuntu 22.04** в **VirtualBox**.

Проверка установки Java и WebKitGTK.

Развертывание Pentaho [Data](http://95.31.0.249/moodle/mod/folder/view.php?id=1500) Integration.

-        **Создать ETL-конвейер**:

Загрузить данные из **CSV-файла**.

Очистить, преобразовать и отфильтровать данные.

Выполнить замену значений.

Выгрузить обработанные данные в **MySQL или PostgreSQL**.

-        **Проверить корректность обработки**:

Выполнить SQL-запросы для проверки результата.

Подготовить отчет с описанием проделанных шагов.

Инструменты и технологии

Для выполнения лабораторной работы используются следующие инструменты:

|  |  |
| --- | --- |
| **Компонент** | **Описание** |
| **Ubuntu 22.04 (.ova)** | Образ операционной системы для VirtualBox 7.0 |
| **VirtualBox 7.0** | Виртуализация среды |
| **Pentaho**[**Data**](http://95.31.0.249/moodle/mod/folder/view.php?id=1500)**Integration 9.4** | ETL-инструмент для работы с данными |
| **MySQL/PostgreSQL** | База данных для хранения обработанных данных |
| **CSV-файлы** | Исходные данные для обработки |
| **Java 11** | Необходима для работы Pentaho |
| **libwebkitgtk-1.0-0** | Библиотека для корректного запуска Spoon |
| **SQL** | Язык запросов для работы с базами данных |

**Формат предоставления отчета по лабораторной работе:**

1. Скачать CSV-датасет с Kaggle.

2. Создать Pentaho ETL-конвейер с фильтрацией, заменой и обработкой данных.

3. Выгрузить данные в MySQL/PostgreSQL.

4. Подготовить отчет с:

   - Описанием процесса;

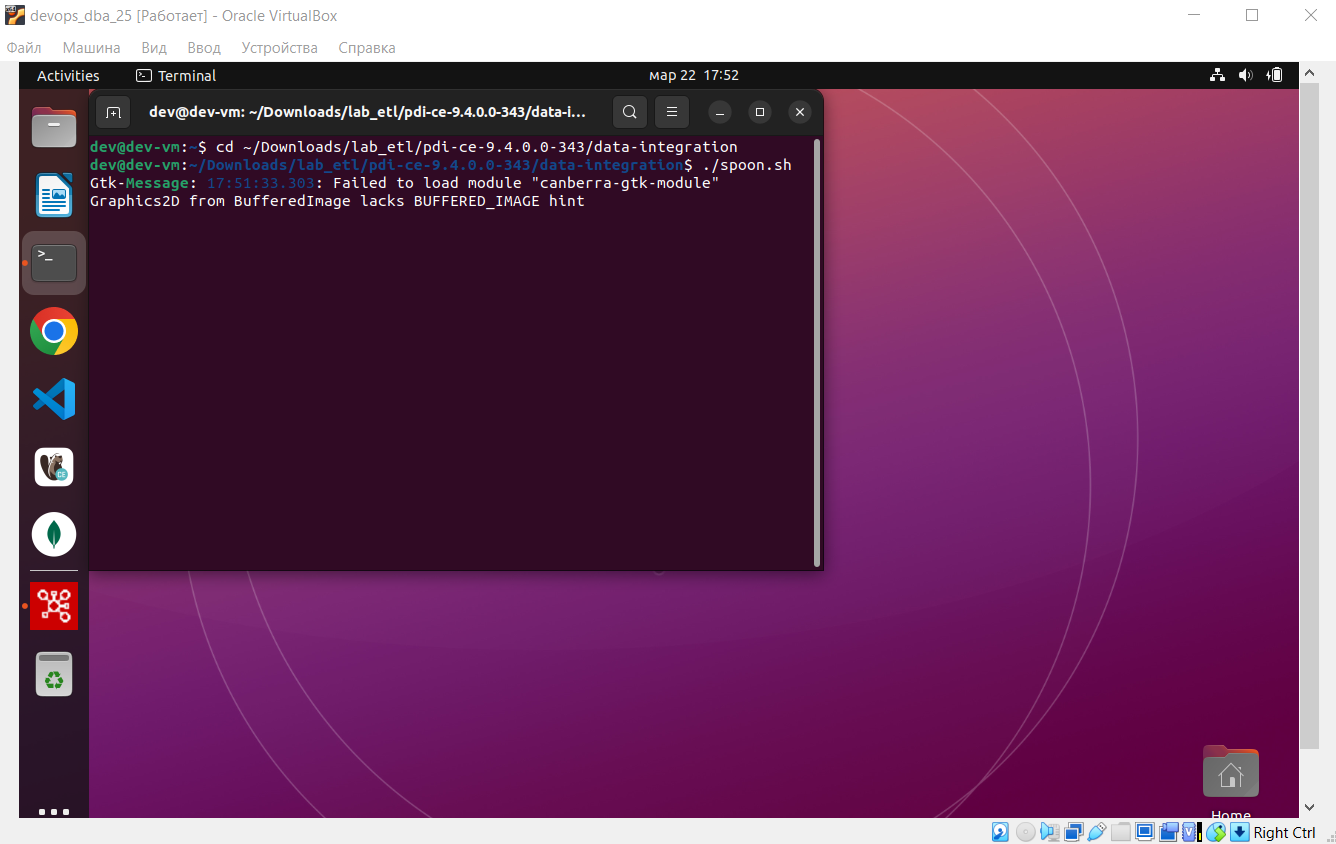
   - Скриншотами Spoon;

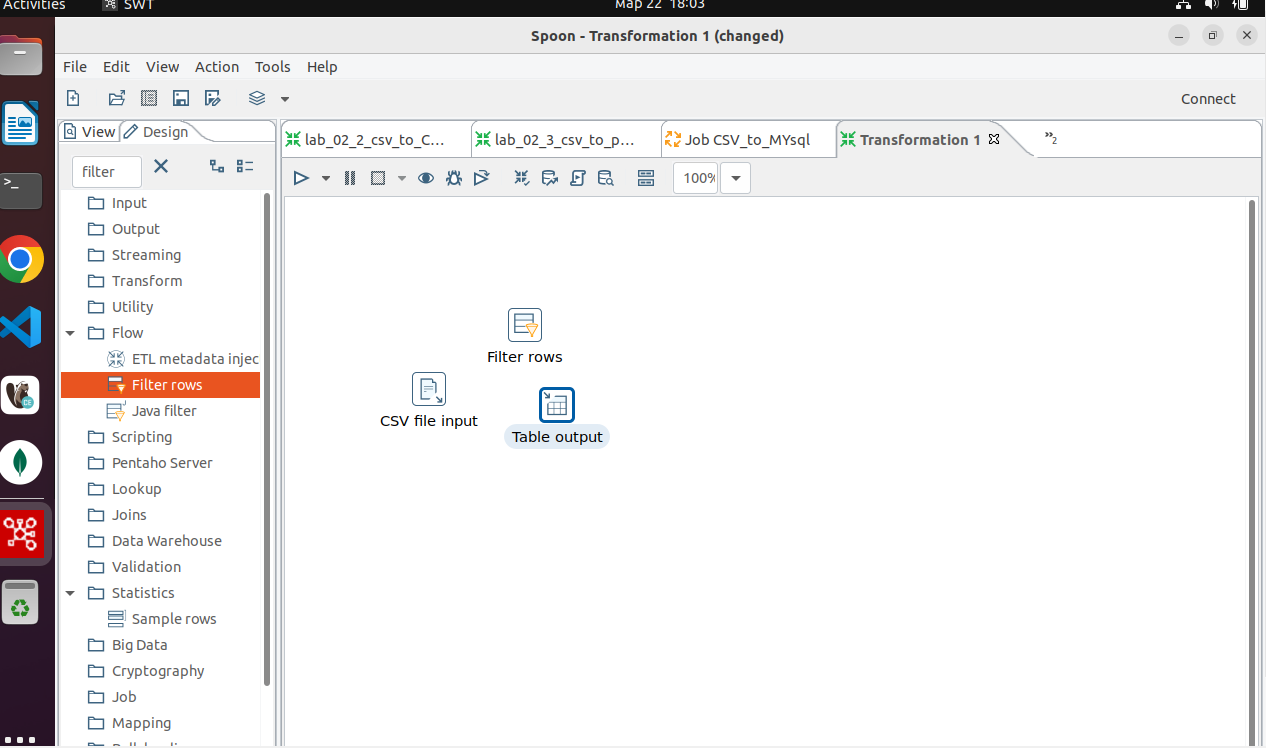
   - SQL-запросами для проверки.

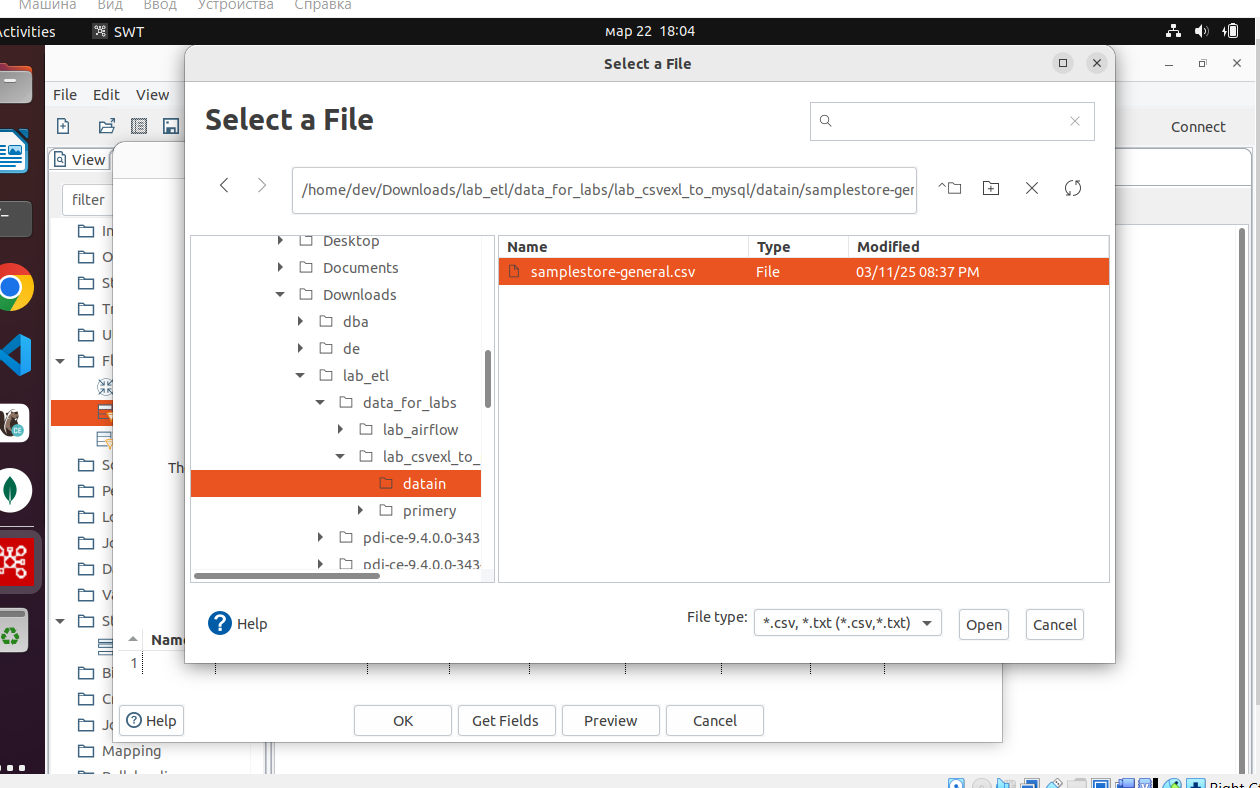
5. Загрузить отчет, CSV-датасет (или ссылку на github) **и lab\_4\_01.ktr** в LMS.

## Основная часть

Pentaho запускаем



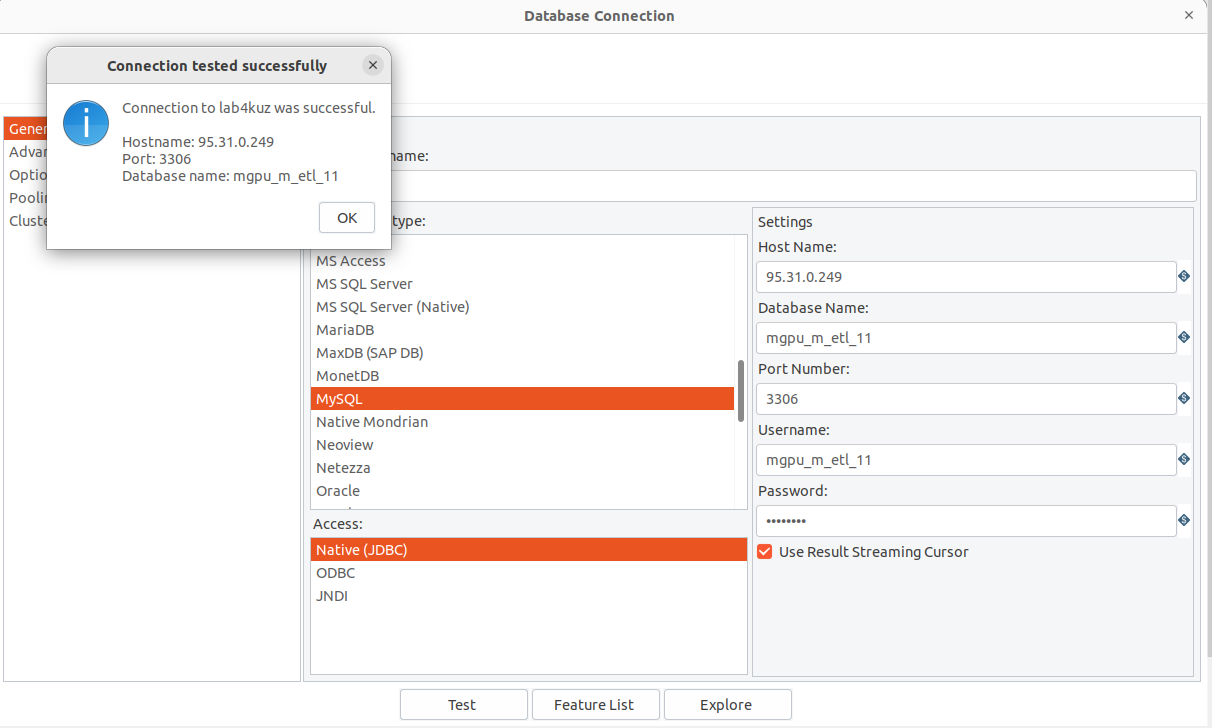




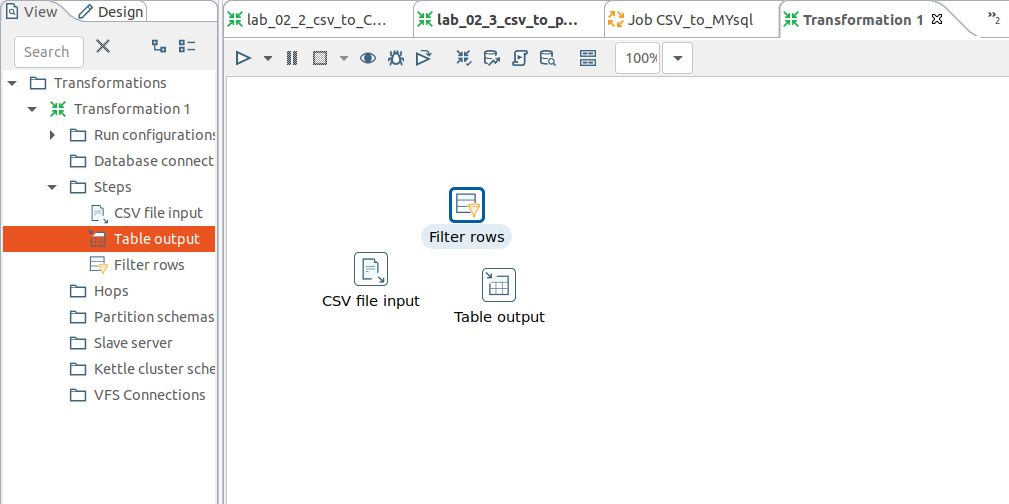
Выбираем файл

Выбираем разделитель

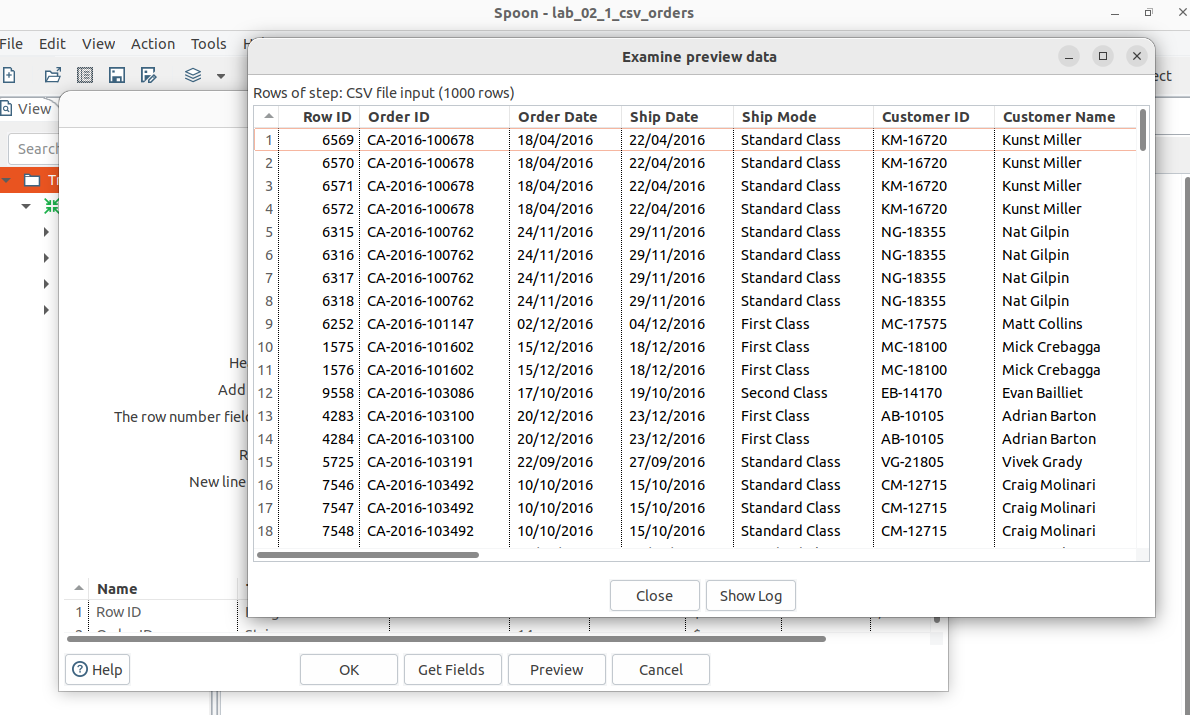
Протестировали и сделали первую трансформацию.

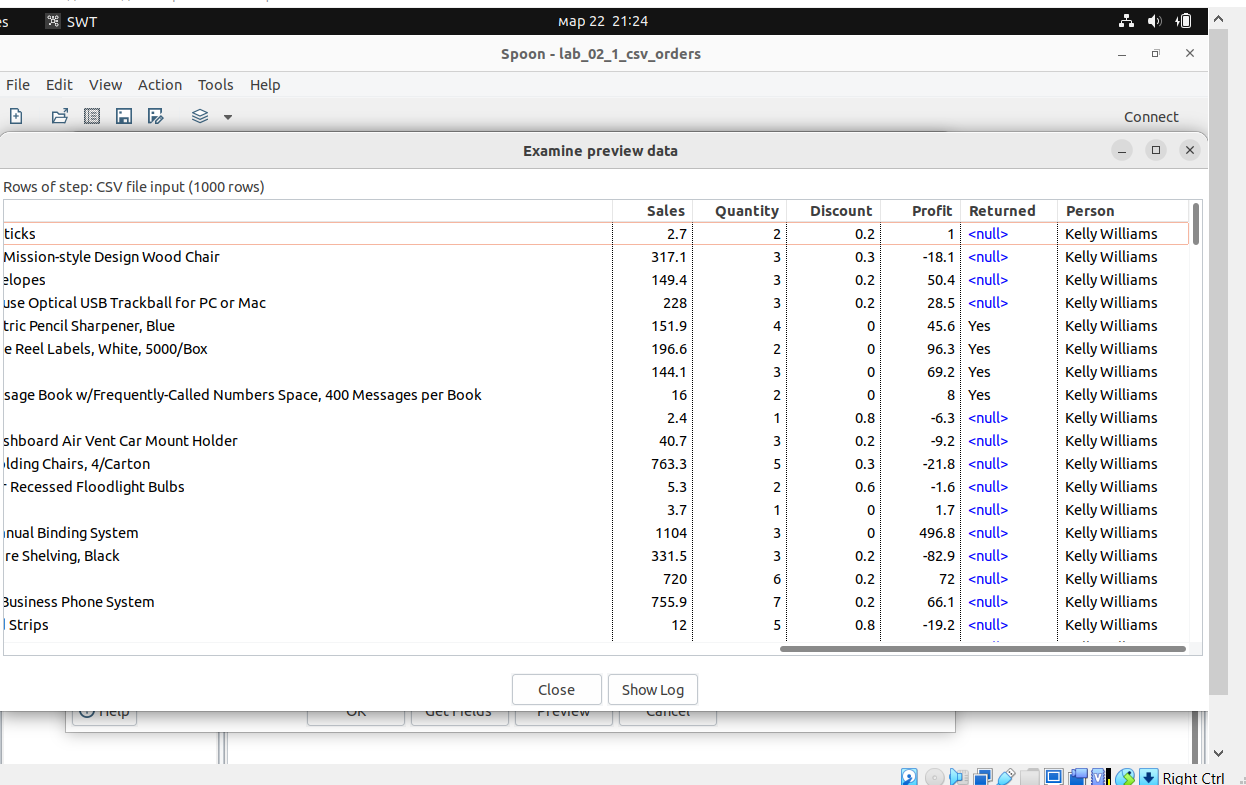


Готово



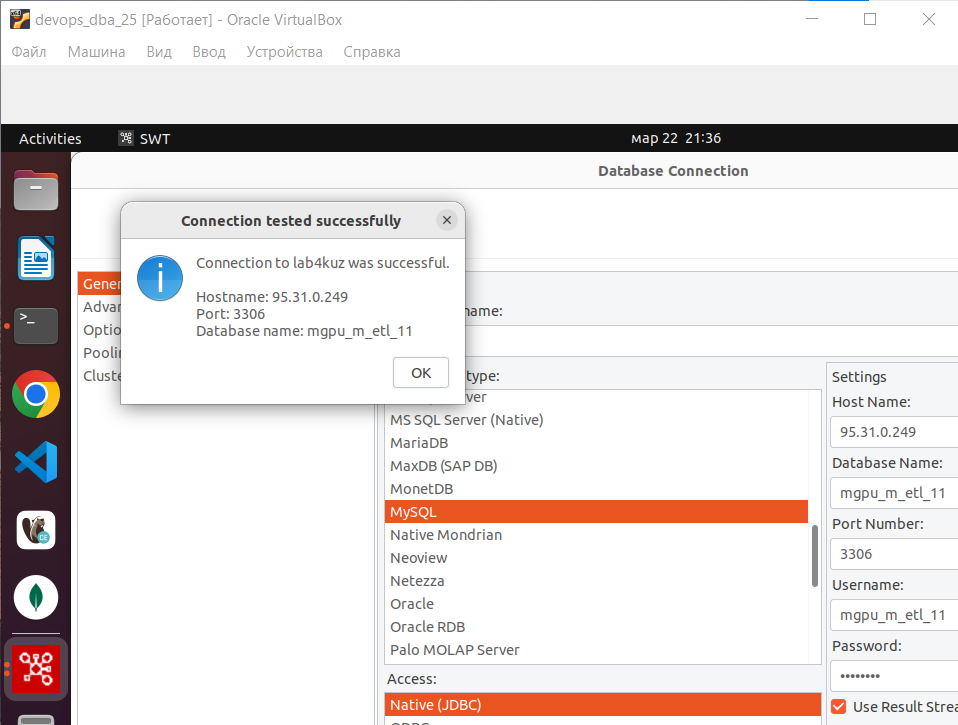
Переходим в заготовку, проверяем зашли ли все данные

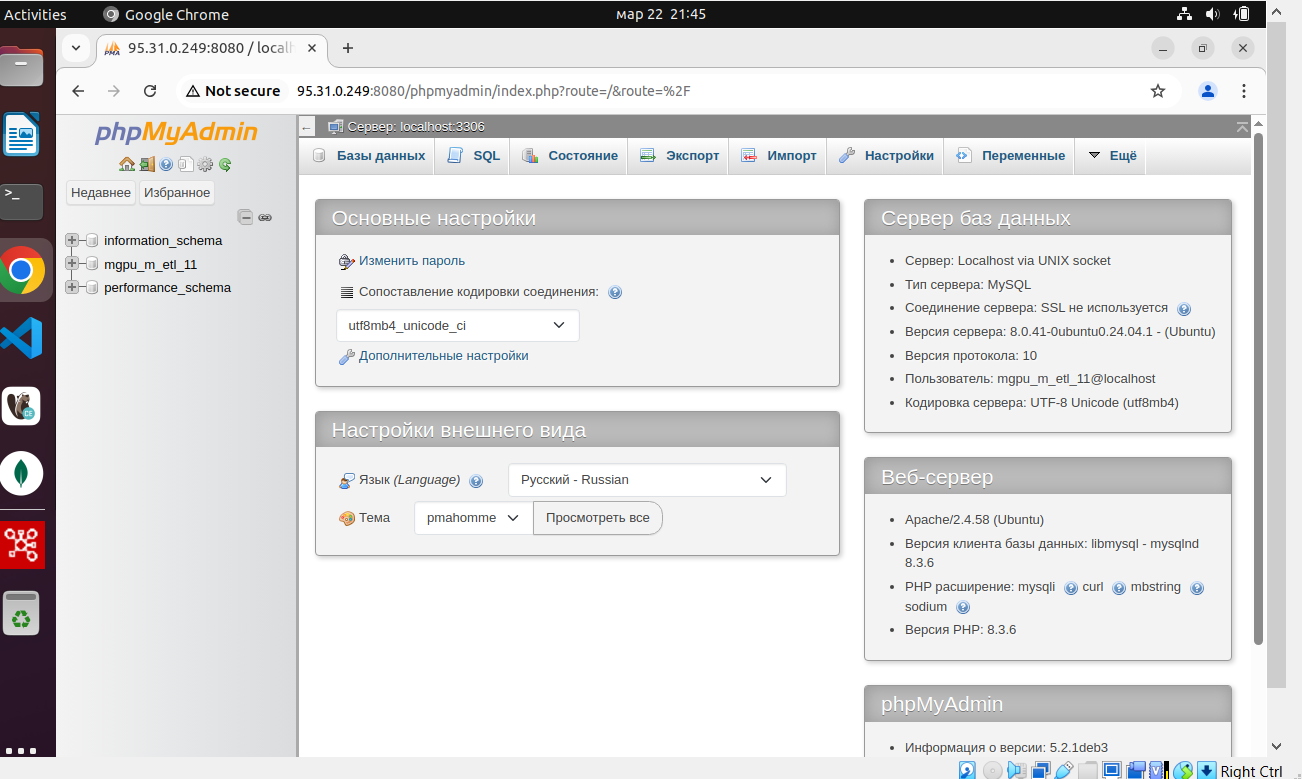


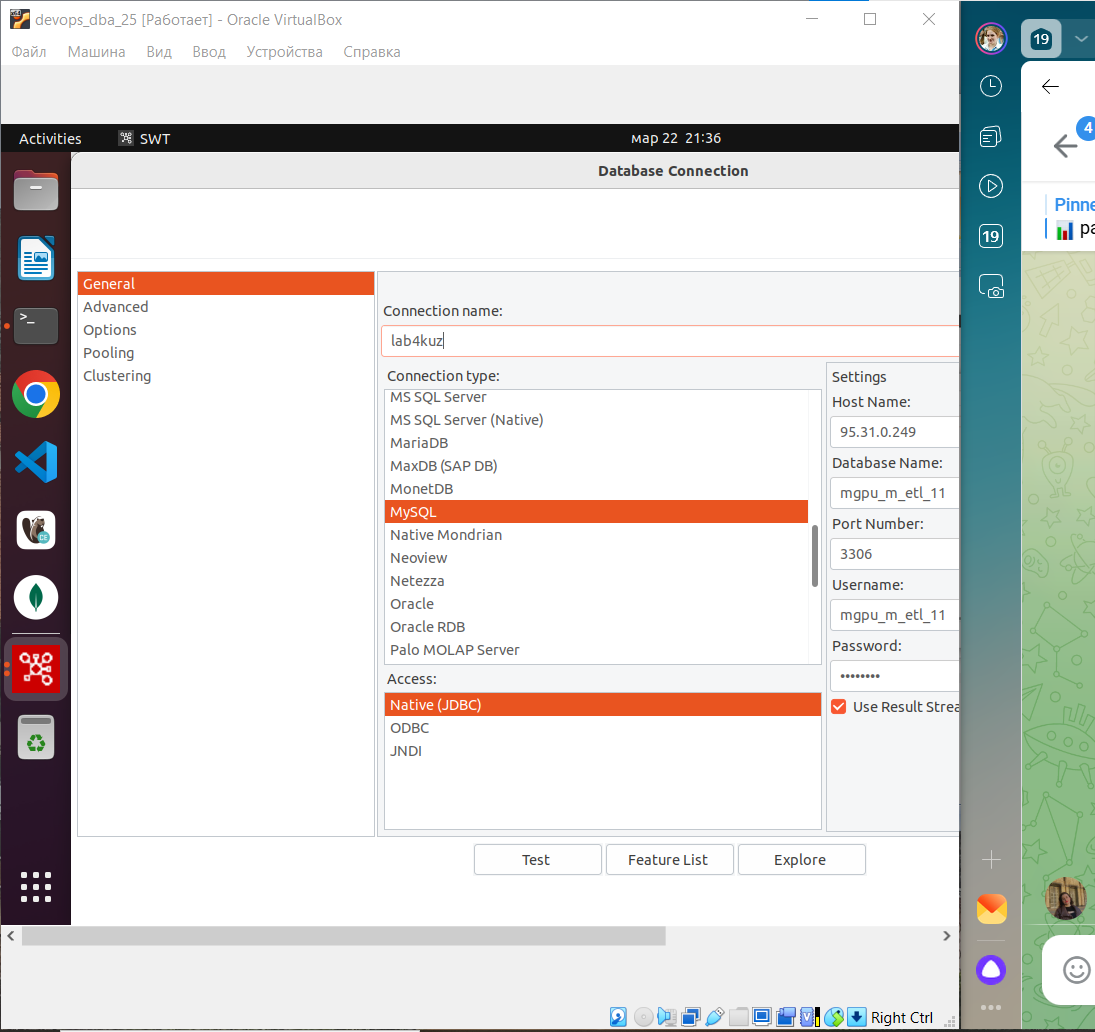


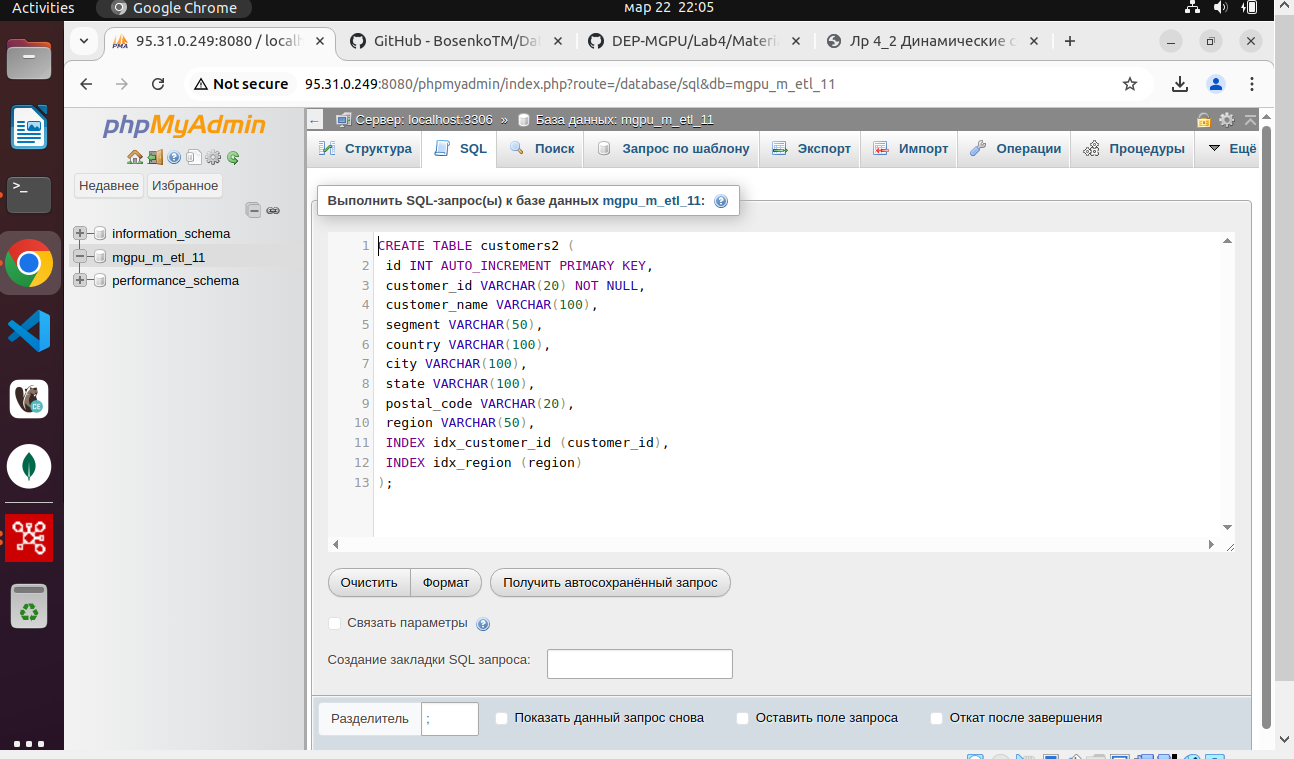
Рассмотрим преобразование значение <null> в 0

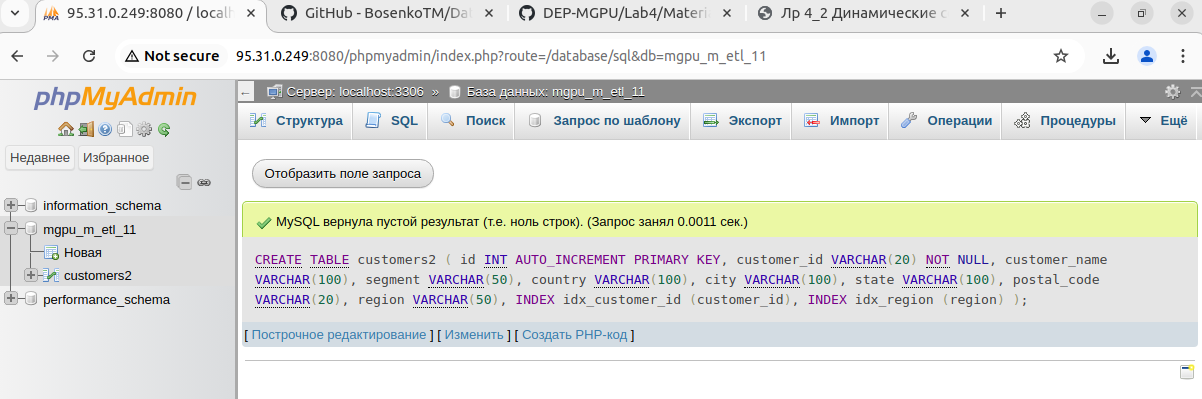
Подключим датасет











# Заключение

В ходе выполнения лабораторной работы были изучены основные принципы работы с ETL-инструментом Pentaho Data Integration (PDI). Была настроена среда для работы с PDI, включая установку Java, WebKitGTK и развертывание Pentaho. Создан ETL-конвейер, который включал загрузку данных из CSV-файла, их очистку, фильтрацию и преобразование. Обработанные данные были выгружены в базу данных MySQL/PostgreSQL.

* Pentaho Data Integration предоставляет удобный интерфейс для создания ETL-процессов.
* Важным этапом является предварительная очистка и трансформация данных для обеспечения их качества.
* Интеграция данных из CSV в базу данных позволяет автоматизировать процессы анализа и отчетности.