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

Создание конвейеров данных

Цель работы: изучение основных принципов работы с ETL-инструментами на примере Pentaho Data 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 Integration, либо проводится установка окружения в ОС Linux.

Задачи

- **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.

- **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 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;

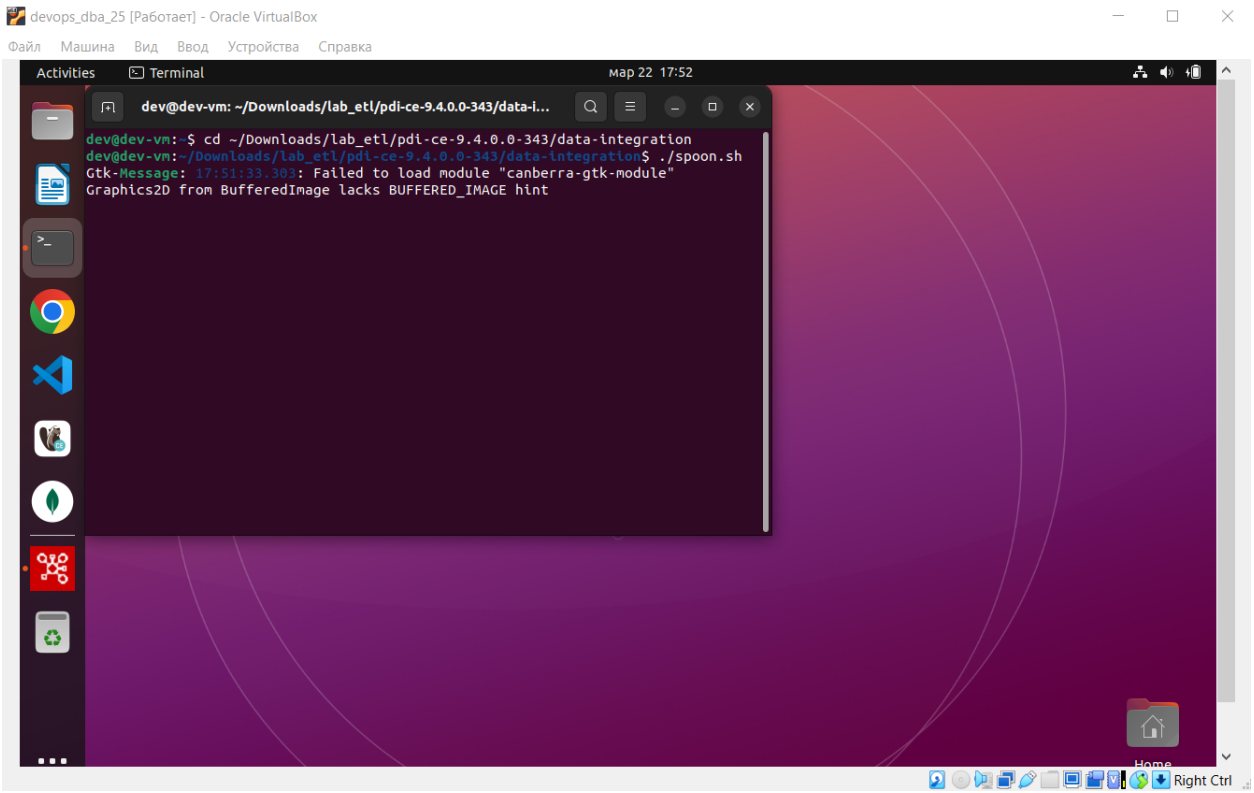
   - Скриншотами Spoon;

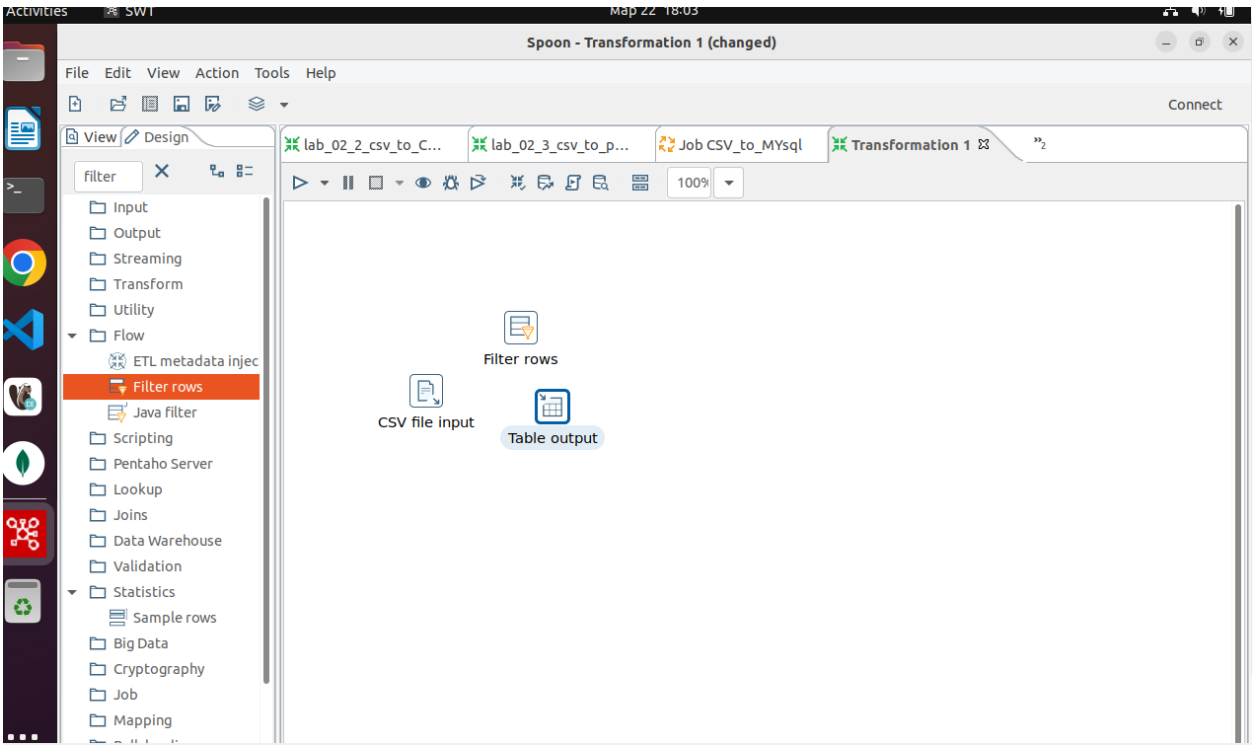
   - SQL-запросами для проверки.

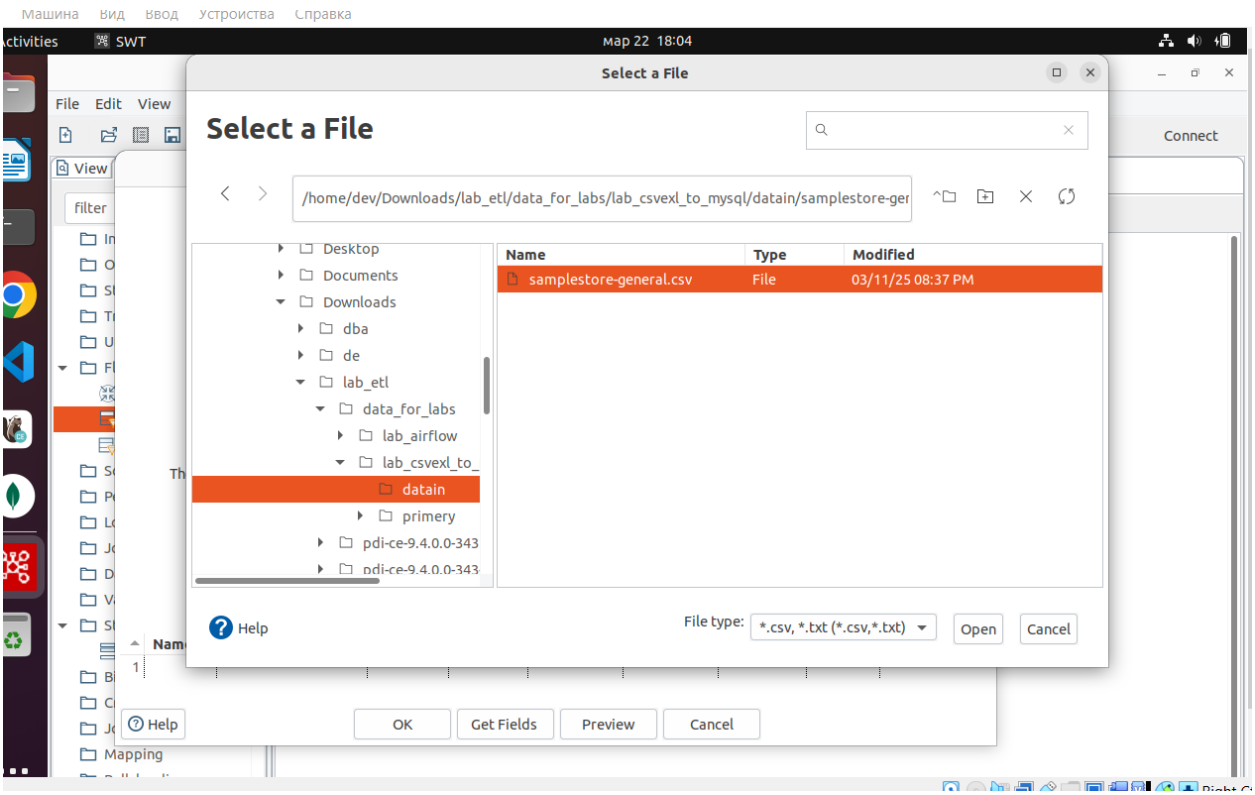
5. Загрузить отчет, CSV-датасет (или ссылку на github) **и lab_4_01.ktr** в LMS.

Основная часть

Pentaho запускаем



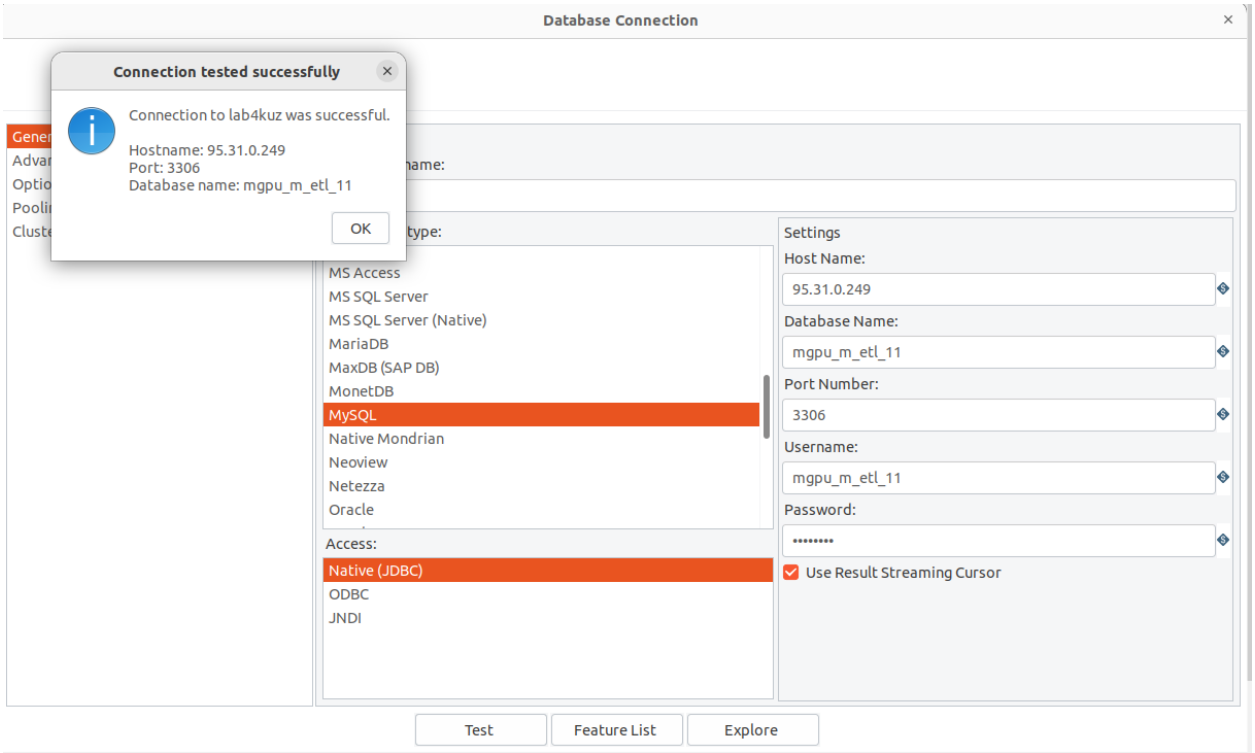




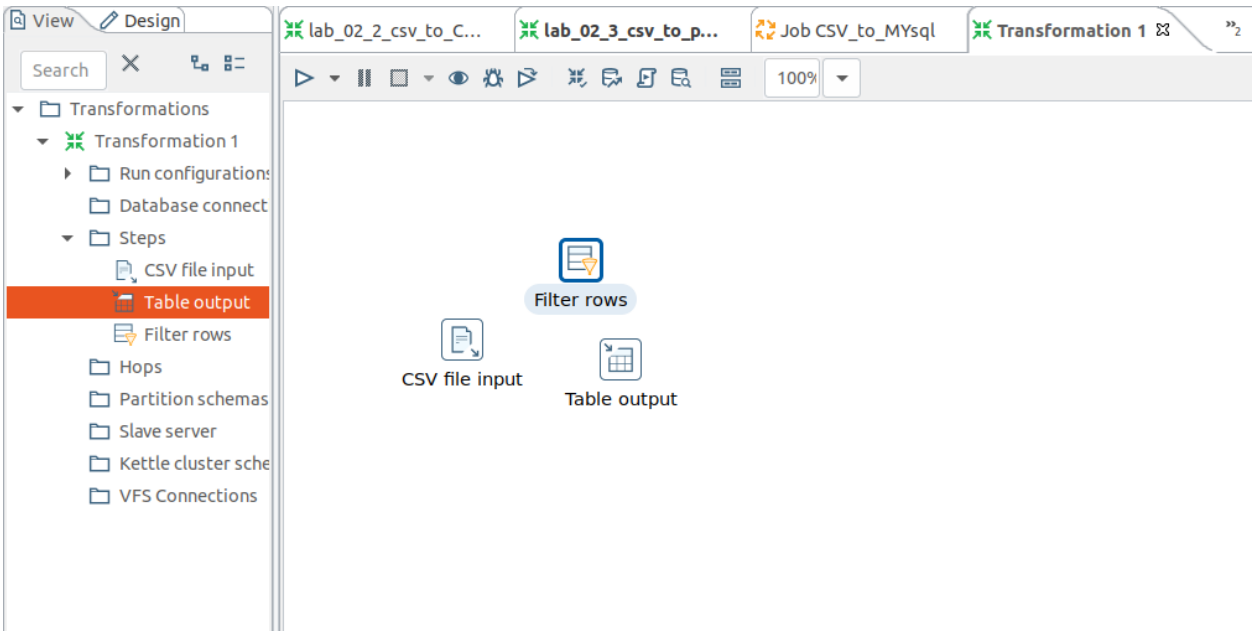
Выбираем файл

Выбираем разделитель

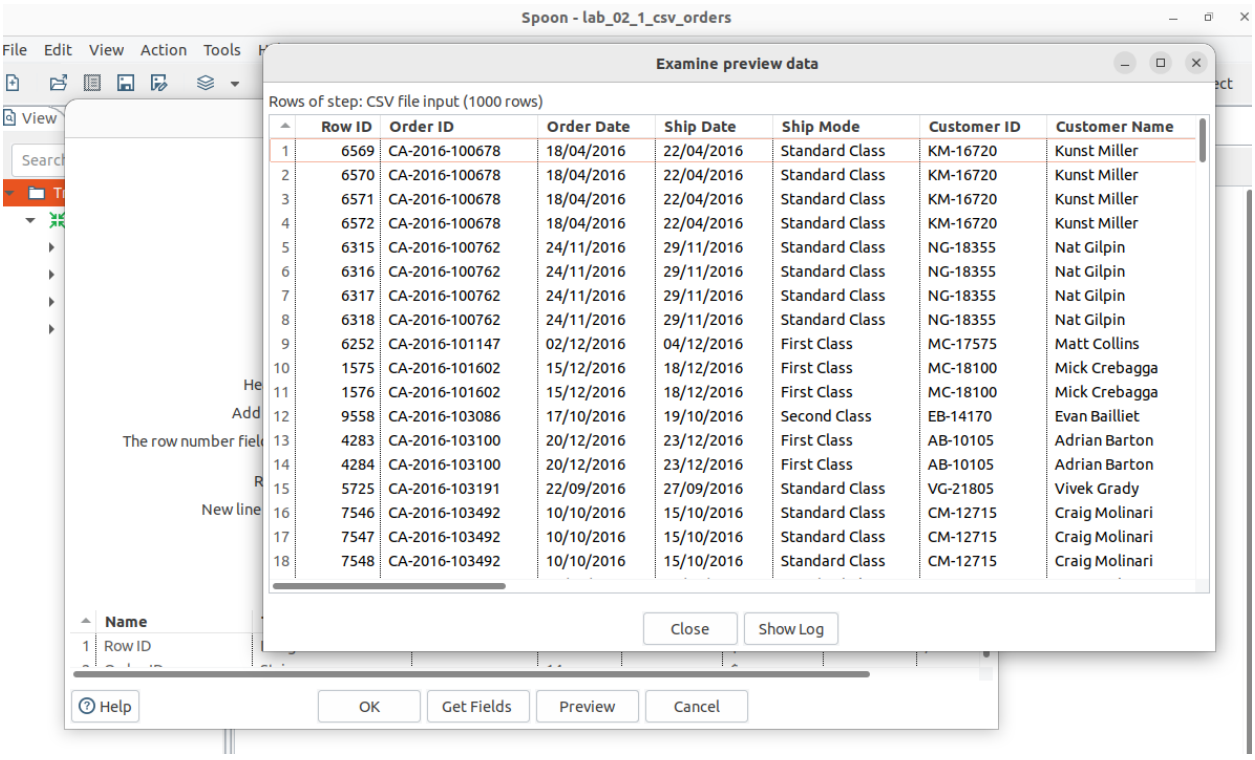
Протестировали и сделали первую трансформацию.

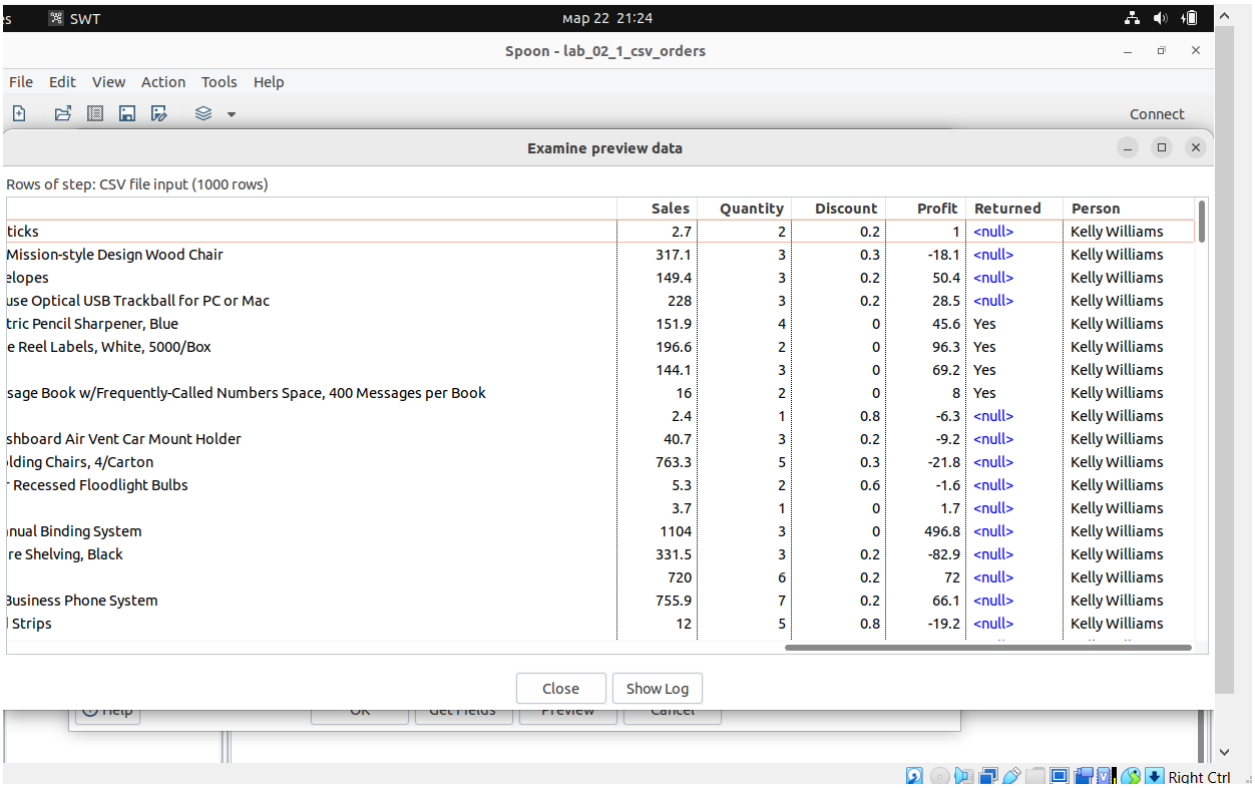


Готово



Переходим в заготовку, проверяем зашли ли все данные

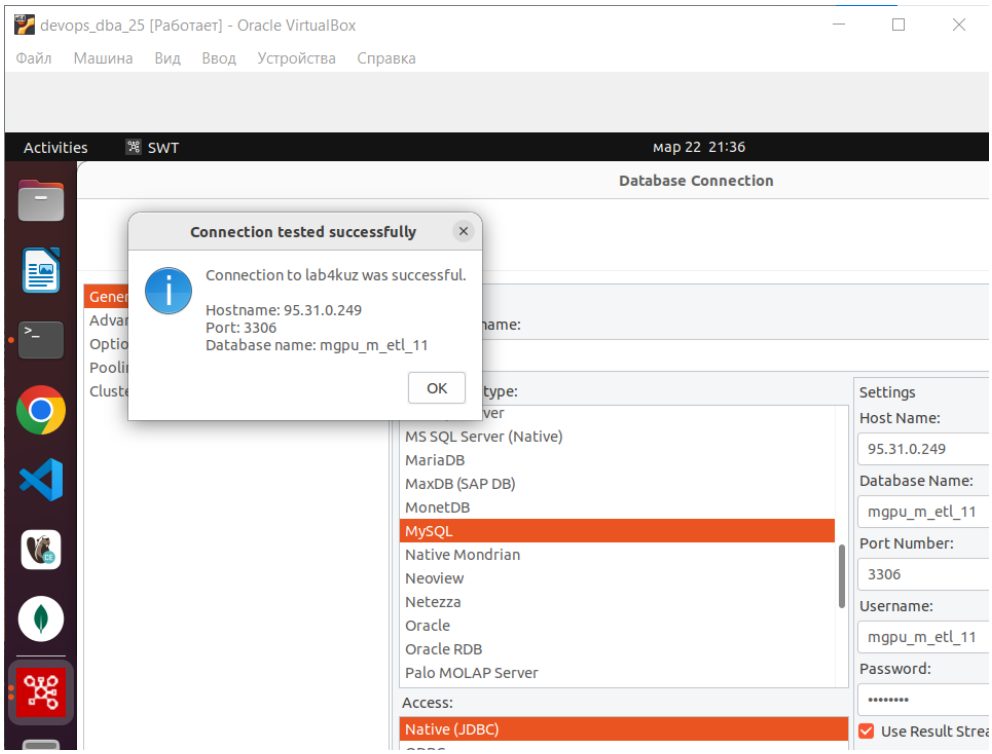


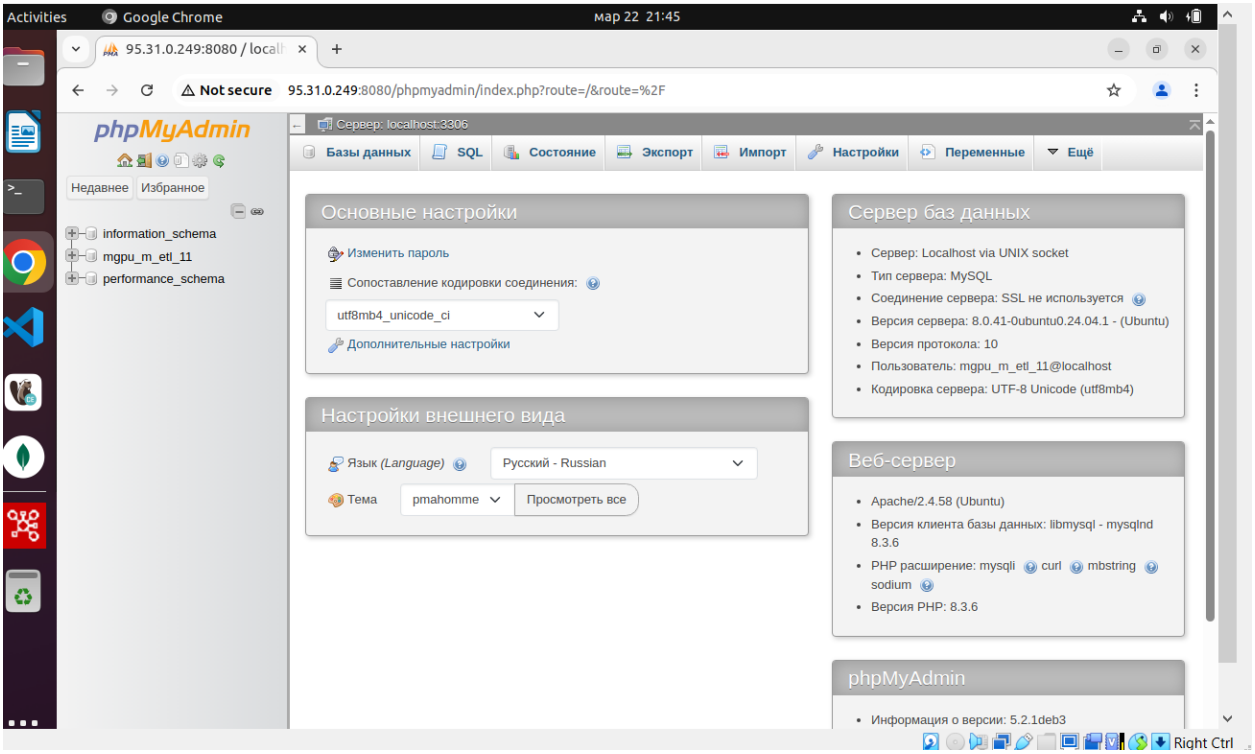


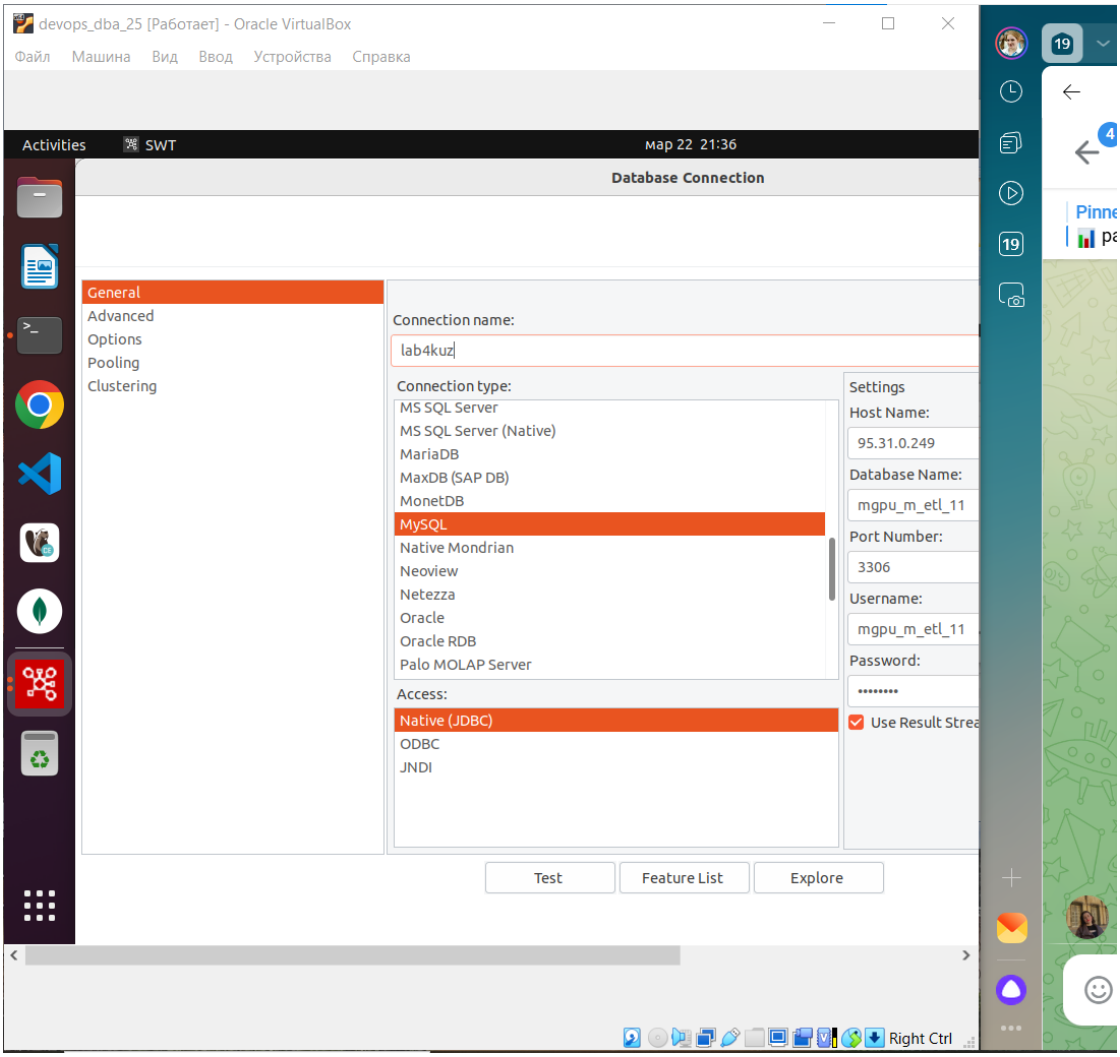
	Sales	Quantity	Discount	Profit	Returned	Person
ticks	2.7	2	0.2	1	<null>	Kelly Williams
Mission-style Design Wood Chair	317.1	3	0.3	-18.1	<null>	Kelly Williams
elopes	149.4	3	0.2	50.4	<null>	Kelly Williams
use Optical USB Trackball for PC or Mac	228	3	0.2	28.5	<null>	Kelly Williams
tric Pencil Sharpener, Blue	151.9	4	0	45.6	Yes	Kelly Williams
e Reel Labels, White, 5000/Box	196.6	2	0	96.3	Yes	Kelly Williams
	144.1	3	0	69.2	Yes	Kelly Williams
sage Book w/Frequently-Called Numbers Space, 400 Messages per Book	16	2	0	8	Yes	Kelly Williams
	2.4	1	0.8	-6.3	<null>	Kelly Williams
shboard Air Vent Car Mount Holder	40.7	3	0.2	-9.2	<null>	Kelly Williams
lding Chairs, 4/Carton	763.3	5	0.3	-21.8	<null>	Kelly Williams
Recessed Floodlight Bulbs	5.3	2	0.6	-1.6	<null>	Kelly Williams
	3.7	1	0	1.7	<null>	Kelly Williams
nual Binding System	1104	3	0	496.8	<null>	Kelly Williams
re Shelving, Black	331.5	3	0.2	-82.9	<null>	Kelly Williams
	720	6	0.2	72	<null>	Kelly Williams
Business Phone System	755.9	7	0.2	66.1	<null>	Kelly Williams
Strips	12	5	0.8	-19.2	<null>	Kelly Williams

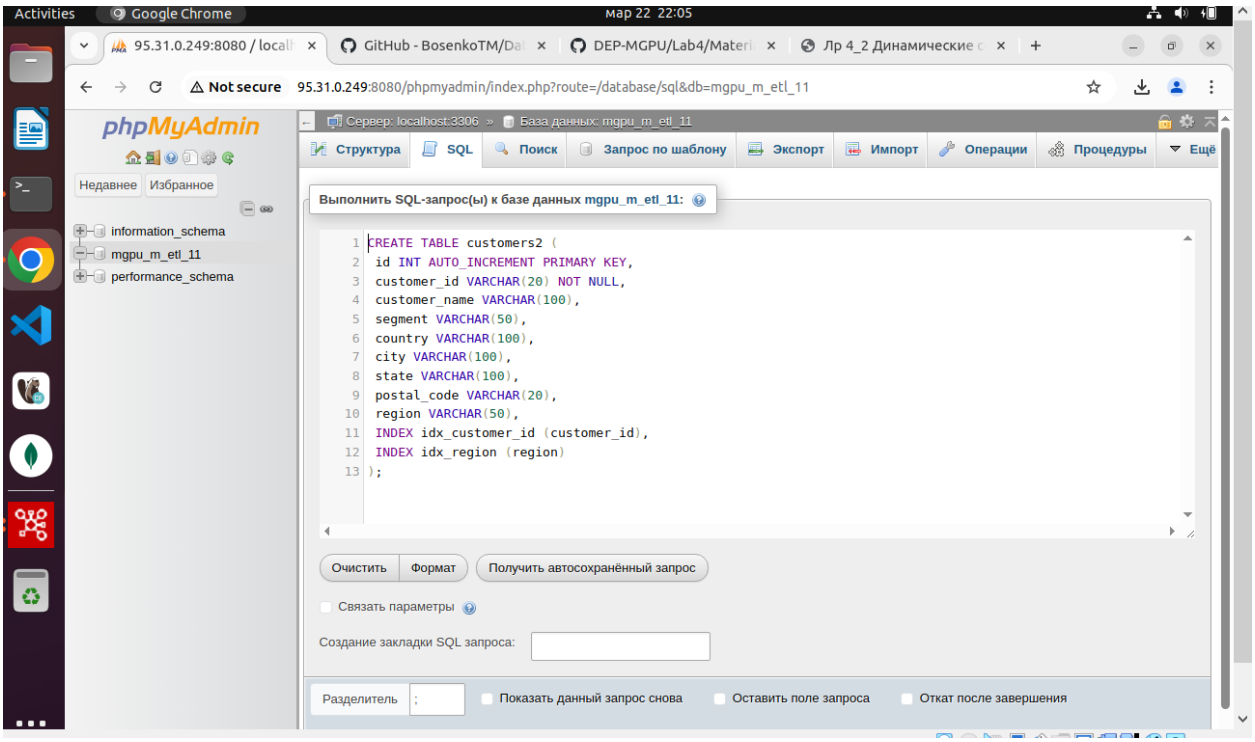
Рассмотрим преобразование значение <null> в 0

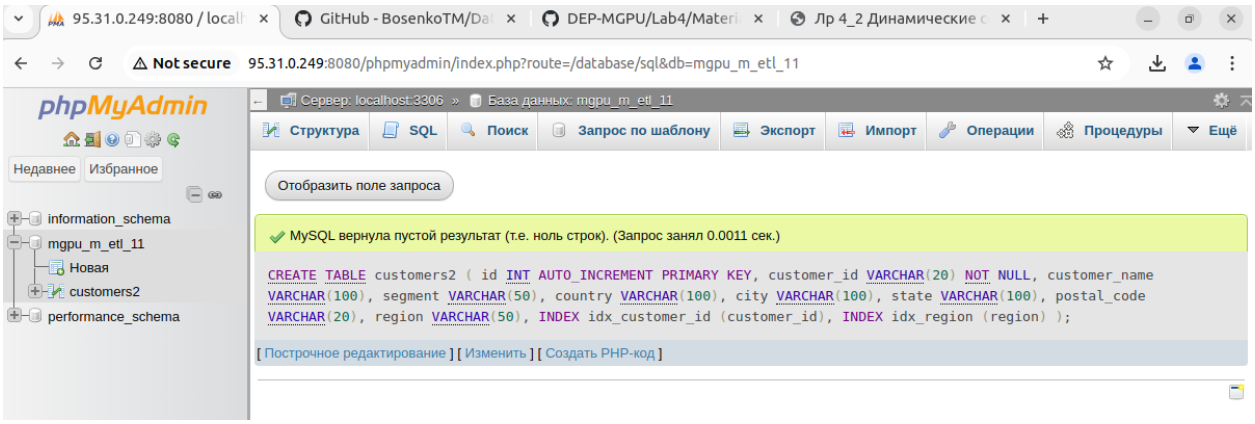
Подключим датасет




Activities
Google Chrome
мар 22 21:45
95.31.0.249:8080 / local
Not secure
95.31.0.249:8080/phpmyadmin/index.php?route=/&route=%2F
phpMyAdmin
Недавнее
Избранное
information_schema
mgpu_m_etl_11
performance_schema
Сервер: localhost:3306
Базы данных
SQL
Состояние
Экспорт
Импорт
Настройки
Переменные
Ещё
Основные настройки
Изменить пароль
Сопоставление кодировки соединения:
utf8mb4_unicode_ci
Дополнительные настройки
Настройки внешнего вида
Язык (Language)
Русский - Russian
Тема
pmahomme
Просмотреть все
Сервер баз данных
Сервер: Localhost via UNIX socket
Тип сервера: MySQL
Соединение сервера: SSL не используется
Версия сервера: 8.0.41-0ubuntu0.24.04.1 - (Ubuntu)
Версия протокола: 10
Пользователь: mgpu_m_etl_11@localhost
Кодировка сервера: UTF-8 Unicode (utf8mb4)
Веб-сервер
Apache/2.4.58 (Ubuntu)
Версия клиента базы данных: libmysql - mysqlnd 8.3.6
PHP расширение: mysqli curl mbstring sodium
Версия PHP: 8.3.6
phpMyAdmin
Информация о версии: 5.2.1deb3
Right Ctrl


devops_dba_25 [Работает] - Oracle VirtualBox
Файл
Машина
Вид
Ввод
Устройства
Справка
Activities
SWT
мар 22 21:36
Database Connection
General
Advanced
Options
Pooling
Clustering
Connection name:
lab4kuz
Connection type:
MS SQL Server
MS SQL Server (Native)
MariaDB
MaxDB (SAP DB)
MonetDB
MySQL
Native Mondrian
Neoview
Netezza
Oracle
Oracle RDB
Palo MOLAP Server
Access:
Native (JDBC)
ODBC
JNDI
Settings
Host Name:
95.31.0.249
Database Name:
mgpu_m_etl_11
Port Number:
3306
Username:
mgpu_m_etl_11
Password:
Use Result Strea
Test
Feature List
Explore
Right Ctrl





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.