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

Смоляков Руслан Игоревич БД-241м

**Лабораторная работа 4 Создание ETL-процесса**

**Вариант 24**

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

**Введение**

***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*](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****.*

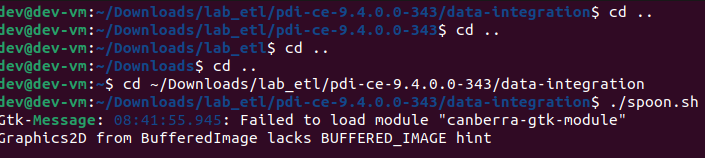
*-****Проверить корректность обработки****:*

*Выполнить SQL-запросы для проверки результата.*

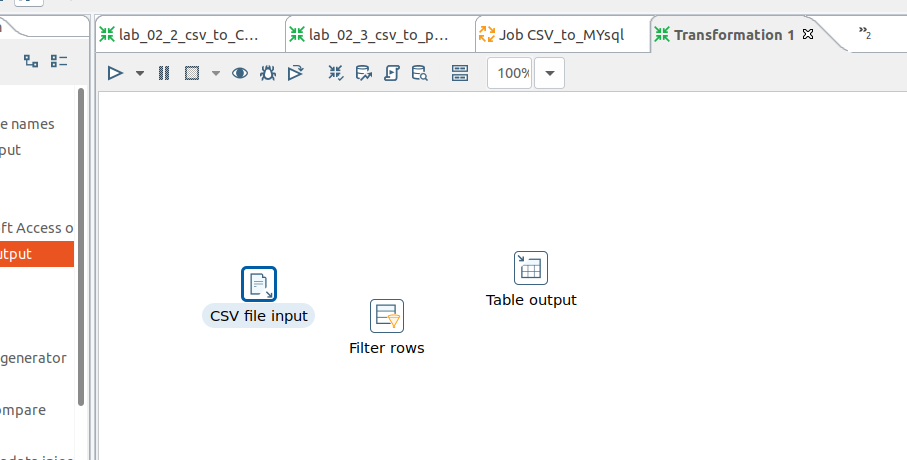
*Подготовить отчет с описанием проделанных шагов.*

**Основная часть**

Выполним команды для открытия Pentaho



Создаем новый проект



File input

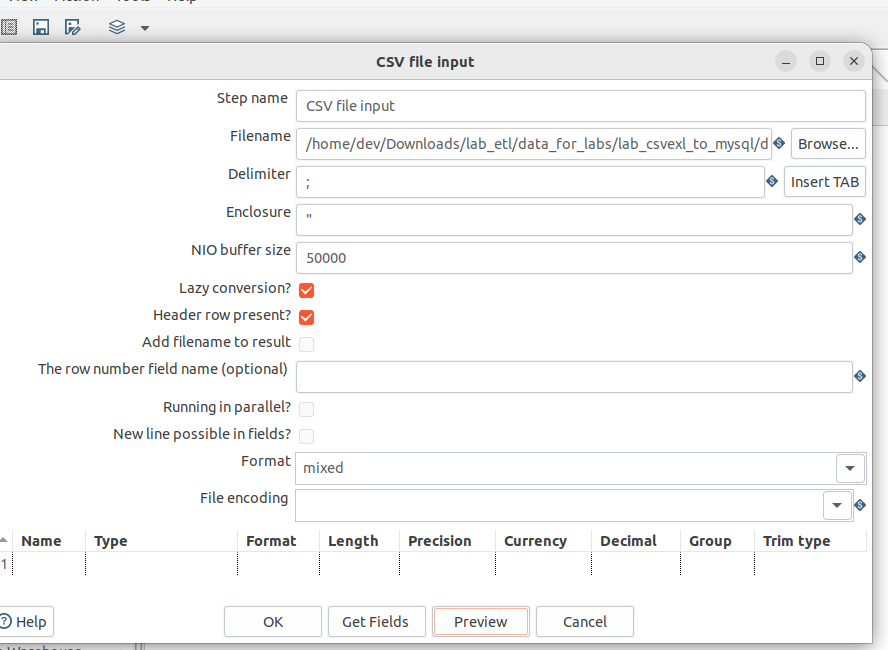
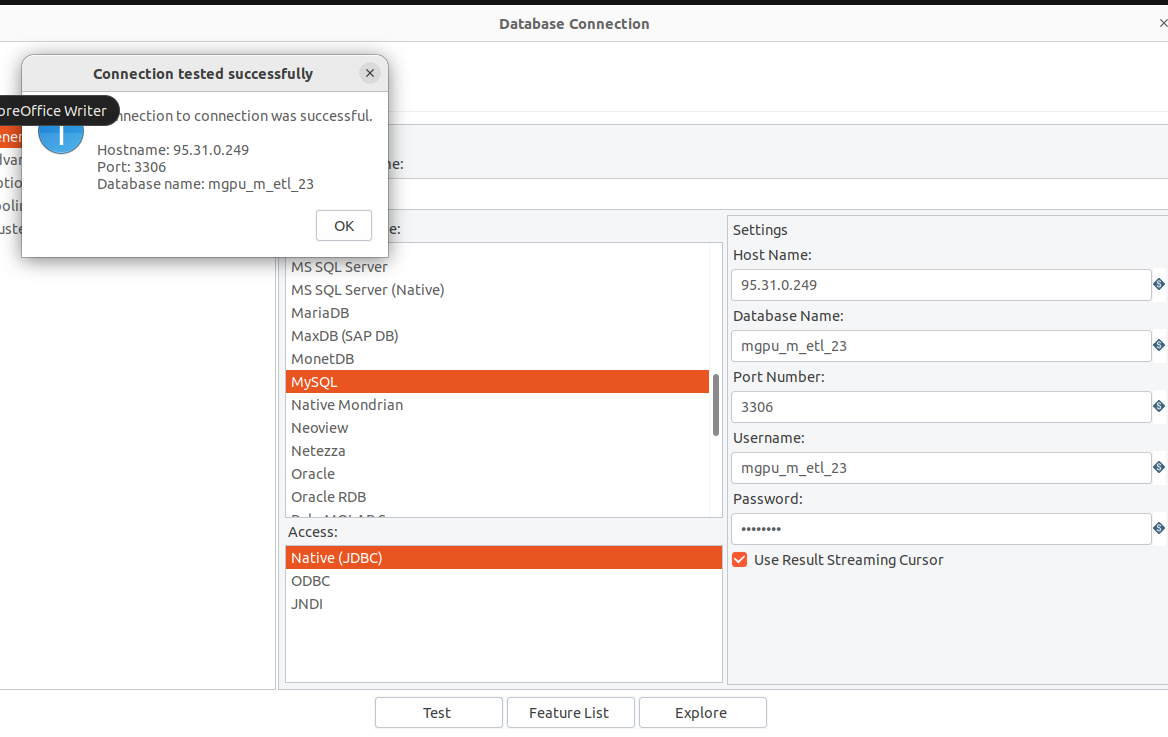


Table output



Успешно подключились

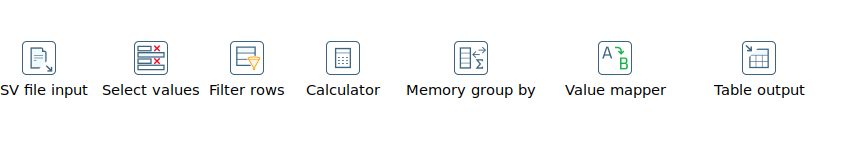
4.1

Вариант 24, всего вариантов 20, взял 4

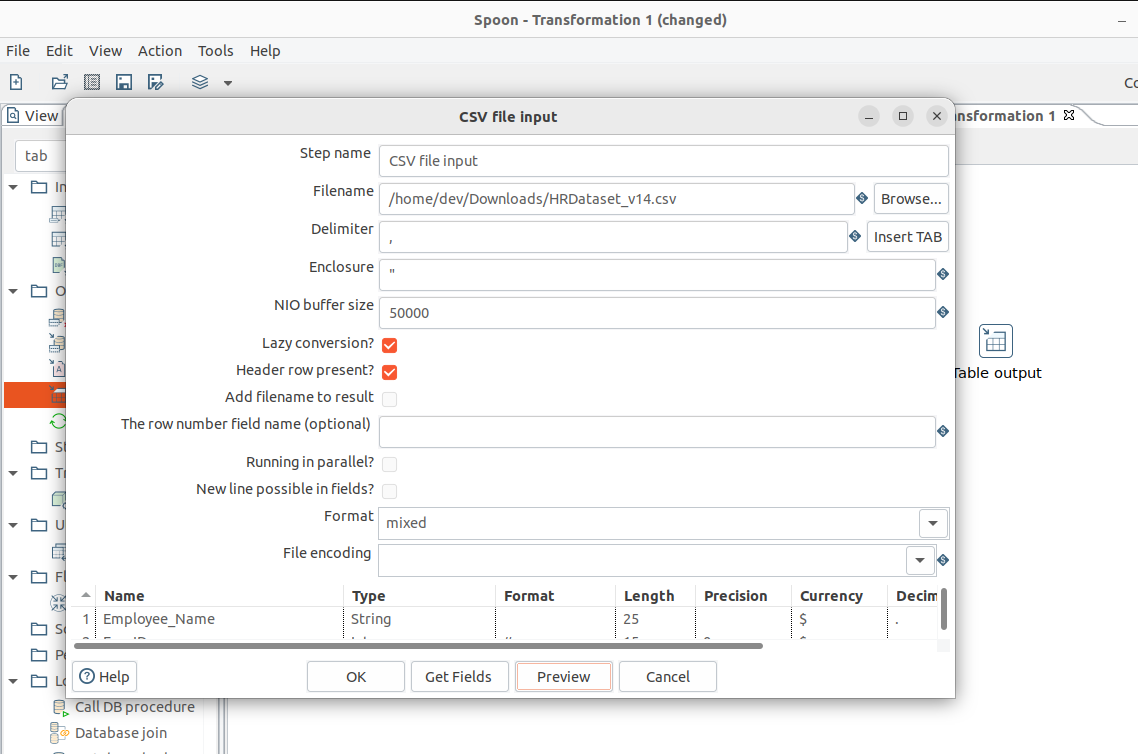
HR-аналитика: анализ данных о сотрудниках, расчет KPI

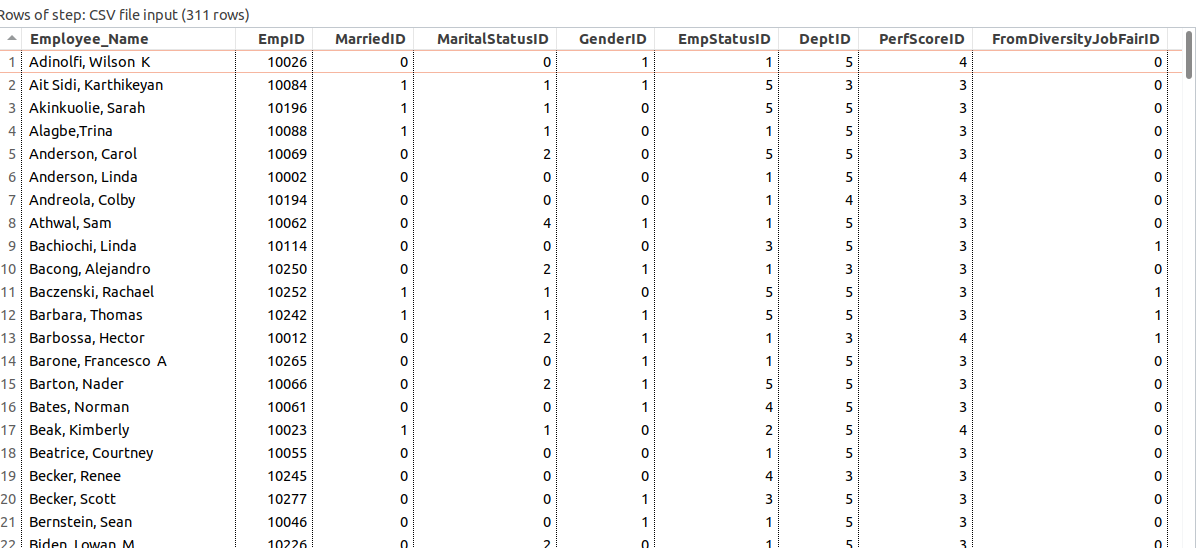
Human Resources Dataset (<https://www.kaggle.com/datasets/rhuebner/human-resources-data-set>)

Трансформация

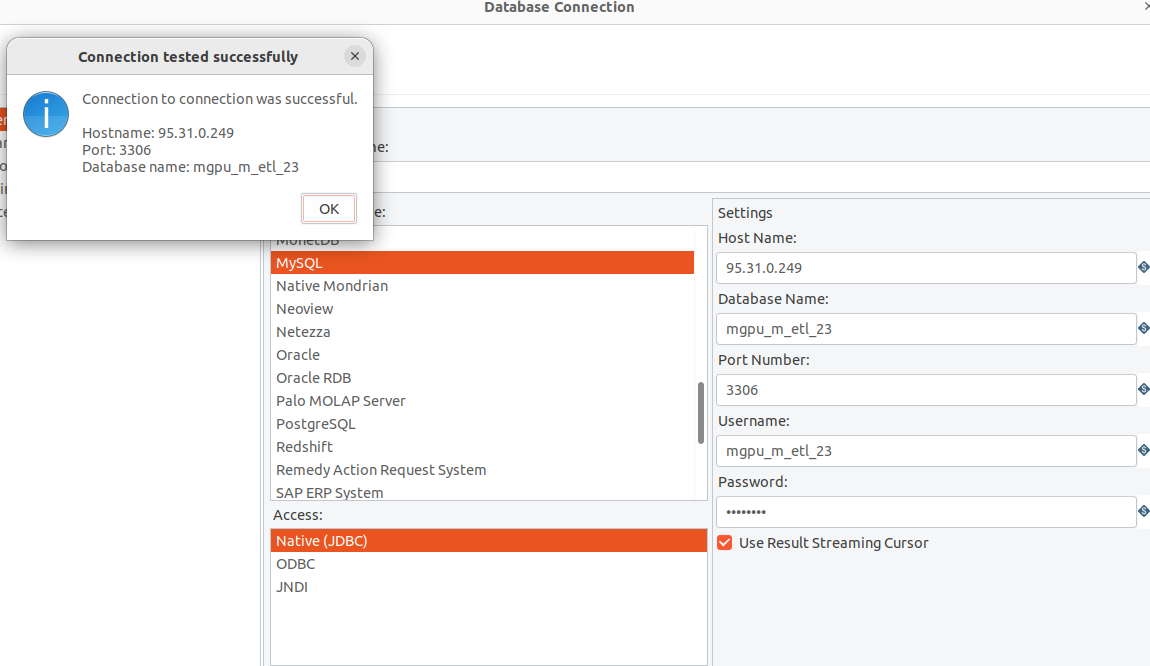


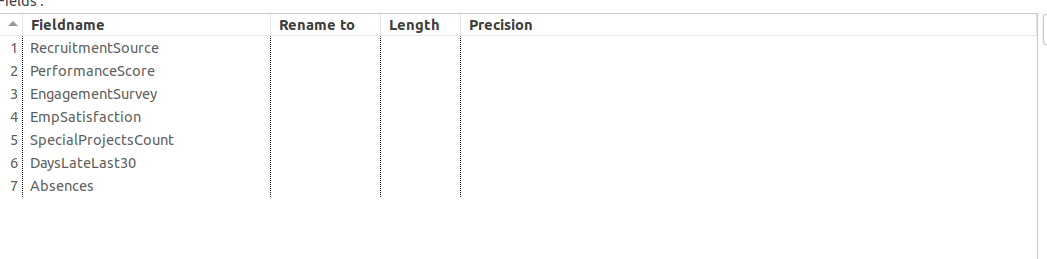
Устанавливаем csv-файл и используем его:



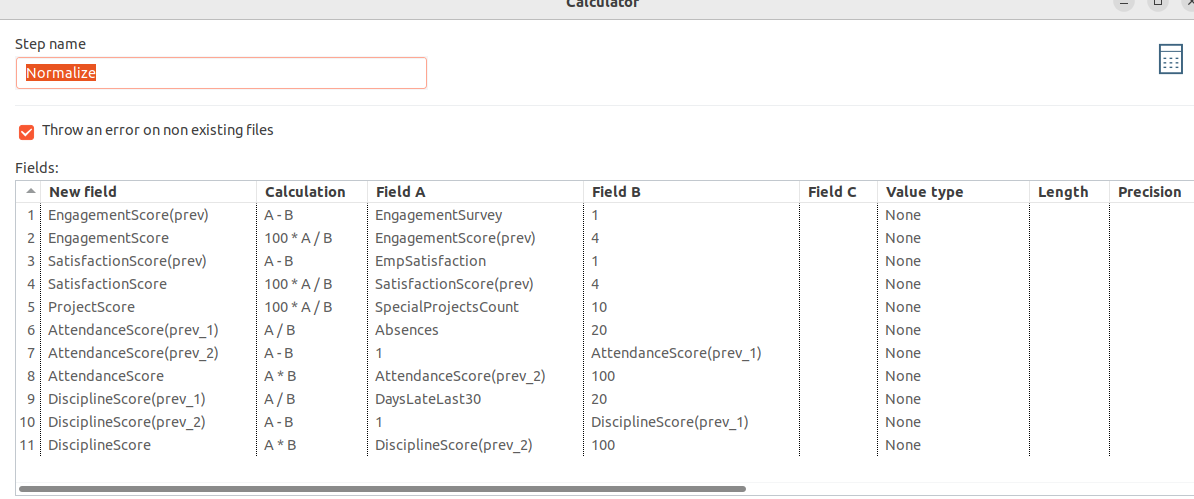
Проверяем, что данные зашли:  


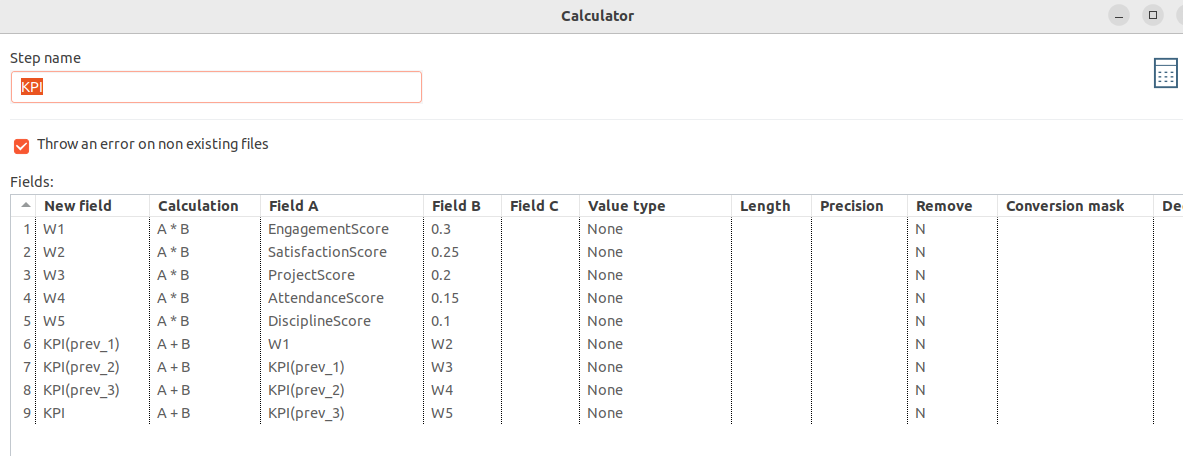
Также успешно подключаемся к phpMyAdmin



Выбираем только нужные строки для Select  


Для расчета KPI выбрал следующие данные:  
Уровень вовлеченности - Engagement Survey Score   
Удовлетворенность работой - Employee Satisfaction   
Количество специальных проектов - Special Projects Count  
Посещаемость – Absences  
Опоздания - DaysLateLast30

Далее нормализуем данные:  


Рассчитываем KPI:  
  
Оставляем только поле KPI