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

Ли Александр Андреевич БД-241м

**Лабораторная работа 4 Создание ETL-процесса**

Направление подготовки/специальность

38.04.05 - Бизнес-информатика

Бизнес-аналитика и большие данные

(очная форма обучения)

Вариант 12

Москва

2024

**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.

Задачи

1.Создать etl-конвейер

2.Проверить корректность обработки

Инструменты для выполнения лабораторной работы

Ubuntu 22.04 (.ova), virtual box 7.0, pentaho data integration 9.4, postgreSql, csv файлы, java 11 , libwebkitgtk-1.0-0

Для начала запускаем pentaho

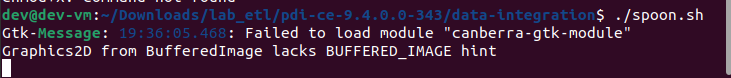


Рис.1 запуск pentaho

Индивидуально задание 1

1. Скачали датасет из Kaggle <https://www.kaggle.com/datasets/arslanali4343/real-estate-dataset/data>

Создали csv file input, прописали путь к файлу и добавили поля, также вручную указали формат полей, так как было много символов после запятой

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

Автоматически созданное описание

Рис.2 создали csv file input

Указываем путь к файлу и разделитель

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

Рис.3 указываем путь к фалу и разделитель

Проверяем через preview импортированные данные.

Изображение выглядит как текст, снимок экрана, число, программное обеспечение

Автоматически созданное описание

Рис.4 проверка импортированных данных

Задаем параметры для select values, так как все значения импортировались как number.

Меняем chas с int на bool.

Zn,rad,tax – так как это целые числа меняем их с number на int

Так же меняем название на более понятные Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

Рис.5 проверка типов данных и переименовываем столбцы

Добавили фильтр замены, так как имеются пустые значения

Если столбец имеет пустые значения, то данные отправляются в логи, иначе данные идут дальше

Изображение выглядит как текст, снимок экрана, программное обеспечение, число

Автоматически созданное описание

Рис.6 задаем параметры для фильтра

Добавили блок value mapper, который заменяет значения 0 и 1 на no и yes, для столбца, который имеет bool

Изображение выглядит как текст, снимок экрана, дисплей, программное обеспечение

Автоматически созданное описание

Рис.7 меняем значения с 1 0 на yes no

Добавили калькулятор, который производит расчет соотношения между стоимостью недвижимости и уровнем преступности и записывает расчеты в новую колонку crime\_per\_pricehouse

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

Рис.7 меняем значения с 1 0 на yes no

После того как наши данные прошли обработку, они отправляются в базу данных, указываем данные для подключения и ниже сверяем столбцы, которые будут записаны в бд.

Изображение выглядит как текст, снимок экрана, программное обеспечение, дисплей

Автоматически созданное описание

Рис.8 экспортируем данные

Общий вид трансформации

Изображение выглядит как текст, снимок экрана, диаграмма, Шрифт

Автоматически созданное описание

Рис.8 трансформация

Создадим базу данных для наших данных

**CREATE TABLE zadanie1 (**

**id INT AUTO\_INCREMENT PRIMARY KEY, -- Добавляем для id, так как нету своего id**

**CRIMINAL FLOAT,**

**ZN INT,**

**INDUS FLOAT,**

**RIVER\_BORDER TINYINT(1) DEFAULT 0, -- 1 = Yes, 0 = No**

**NOX FLOAT,**

**AVG\_ROOM FLOAT,**

**AGE FLOAT,**

**DIS FLOAT,**

**MAGISTRAL TINYINT(1) DEFAULT 0,**

**TAX INT,**

**PTRATIO FLOAT,**

**B FLOAT,**

**LSTAT FLOAT,**

**MED\_PRICE\_HOUSE float,**

**criminal\_per\_price FLOAT**

**);**

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

Автоматически созданное описание**

Рис.8 заполняем данными созданную таблицу

Заключение

В ходе проделанной лабораторной работы, был построен etl-процесс, исходные данные были преобразованы и загружены в базу данных phpMyAdmin.