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

ДИСЦИПЛИНА:

Проектный практикум по разработке ETL-решений

**Лабораторная работа 1**

**Установка и настройка ETL-инструмента.**

**Создание конвейеров данных**

Выполнил(а): Шведова С.С., группа: АДЭУ-211

Преподаватель: Босенко Т.М.

Москва

2025

**Цель работы**: изучение основных принципов работы с ETL-инструментами на примере Pentaho Data Integration (PDI), настройка конвейера обработки данных, фильтрация и замена значений в Excelфайле, а также выгрузка обработанных данных в базу данных MySQL/PostgreSQL.

**Задачи**

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

* Запуск виртуальной машины с Ubuntu 22.04 в VirtualBox.
* Проверка установки Java и WebKitGTK.
* Pentaho Data Integration.

− Создать ETL-конвейер:

* Очистить, преобразовать и отфильтровать данные. Выполнить замену значений.
* Загрузить данные из CSV-файла.
* Выгрузить обработанные данные в MySQL или PostgreSQL.

− Проверить корректность обработки:

* Выполнить SQL-запросы для проверки результата.
* Подготовить отчет с описанием проделанных шагов.

**Вариант 13. Анализ социальных медиа: обработка данных активности.**

Для начала надо запустить Pentaho Spoon (рисунок 1).

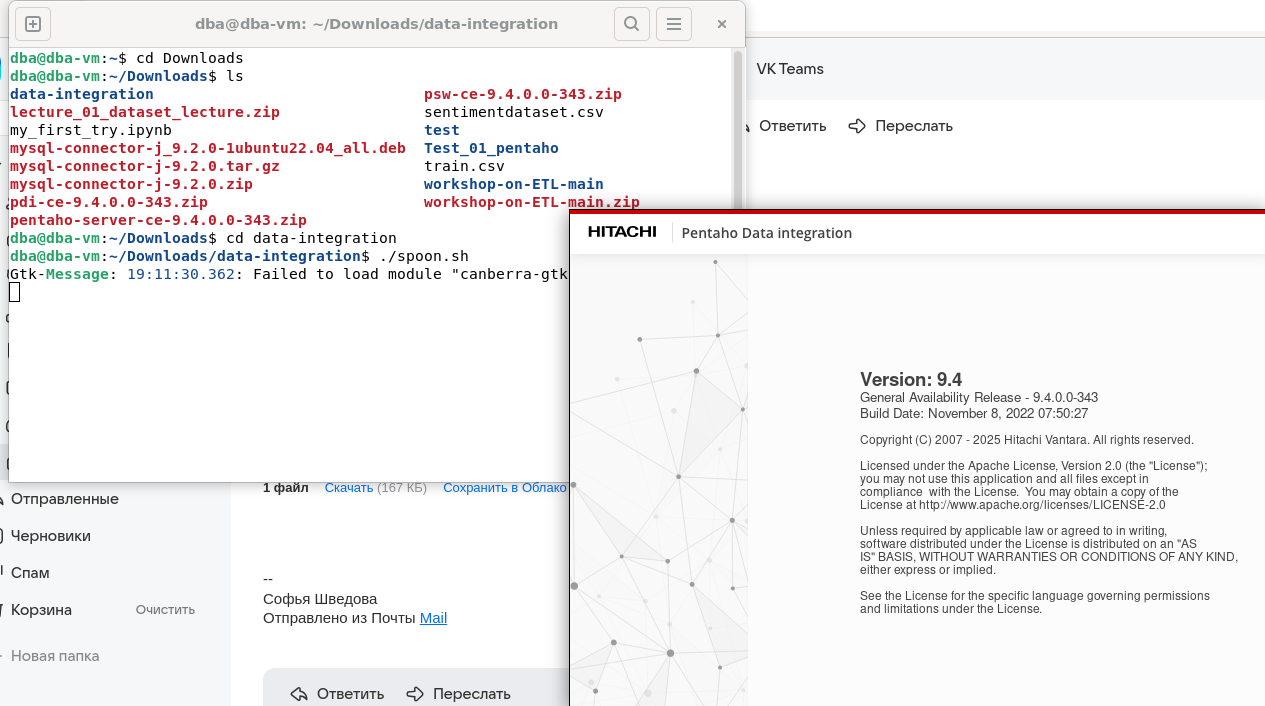


Рисунок 1. Запуск Pentaho Spoon

Затем нужно в PhpMyAdmin создать таблицу Social\_Media\_Sentiments, где будут 4 столбца: Likes, Platform, Sentiment, Retweets, куда будут загружаться данные из Pentaho (рисунок 2).

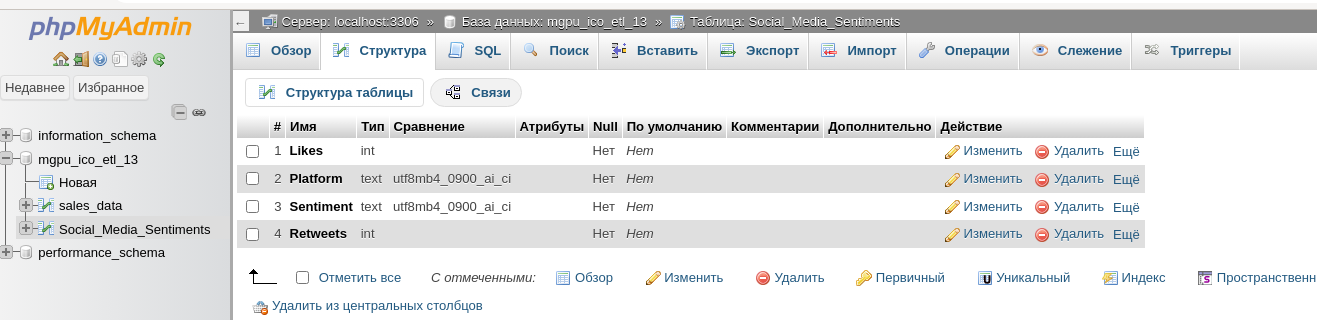


Рисунок 2. Создание таблицы Social\_Media\_Sentiments

Далее надо загрузить CSV файл в Pentaho (рисунок 3).

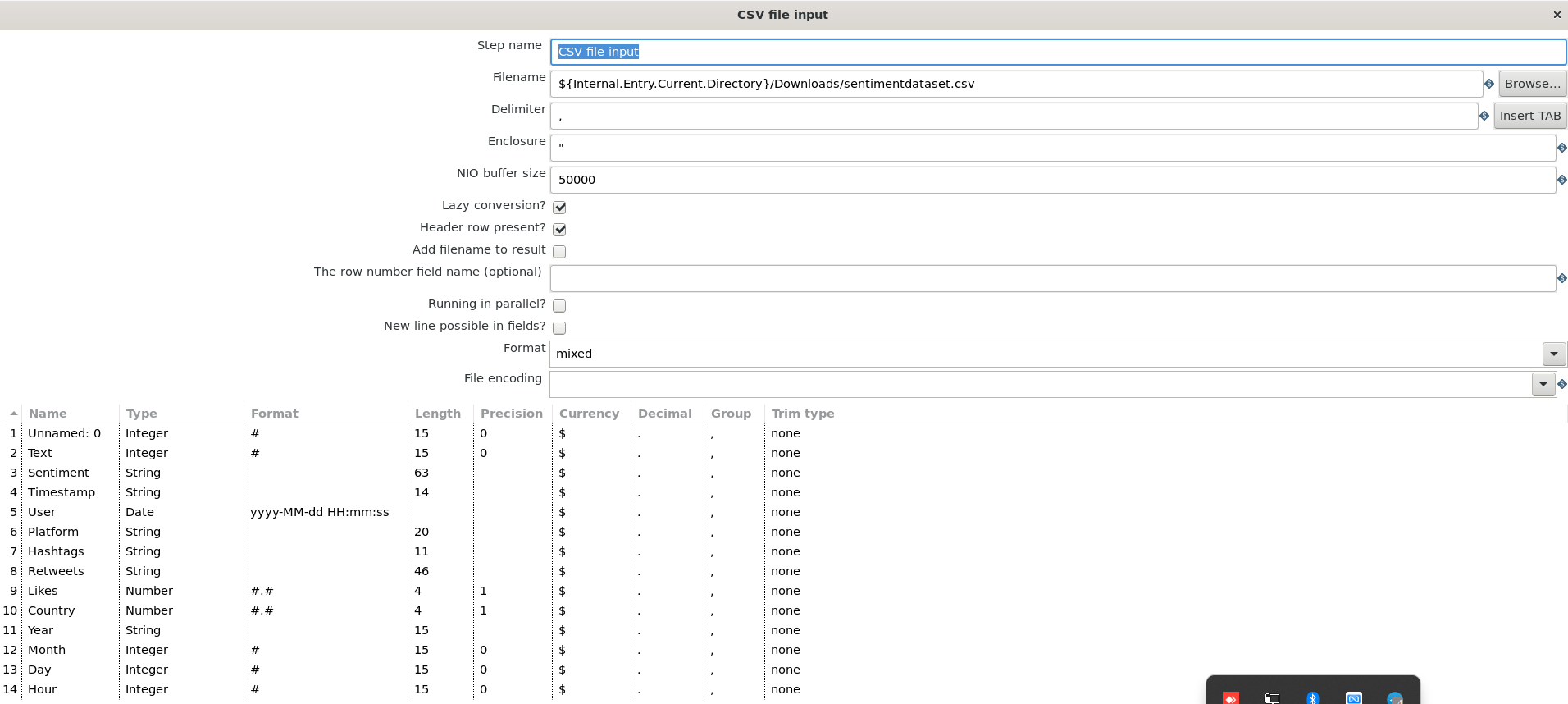
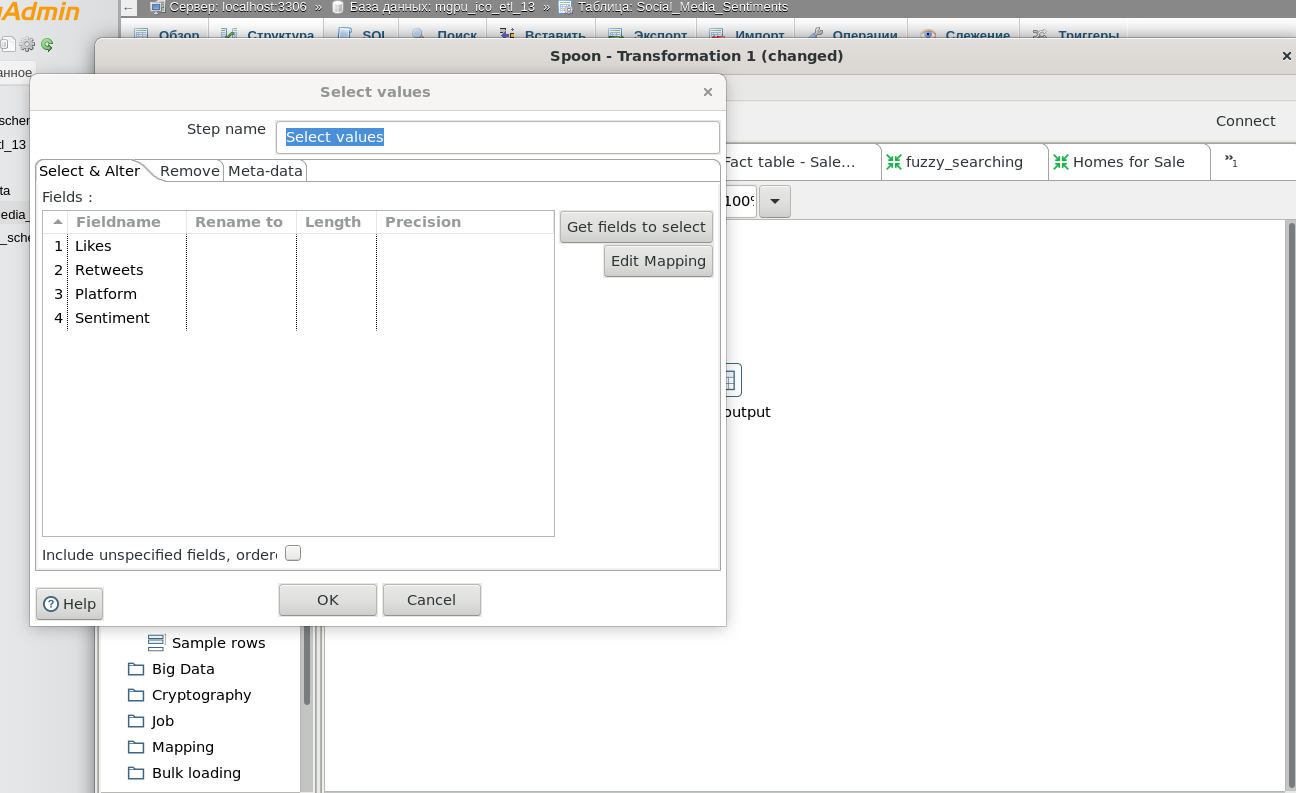


Рисунок 3. Загрузка CSV файла в Pentaho

Теперь надо выбрать столбцы, а именно Likes, Platform, Sentiment, Retweets (рисунок 4).

  
Рисунок 4. Выбор определенных столбцов

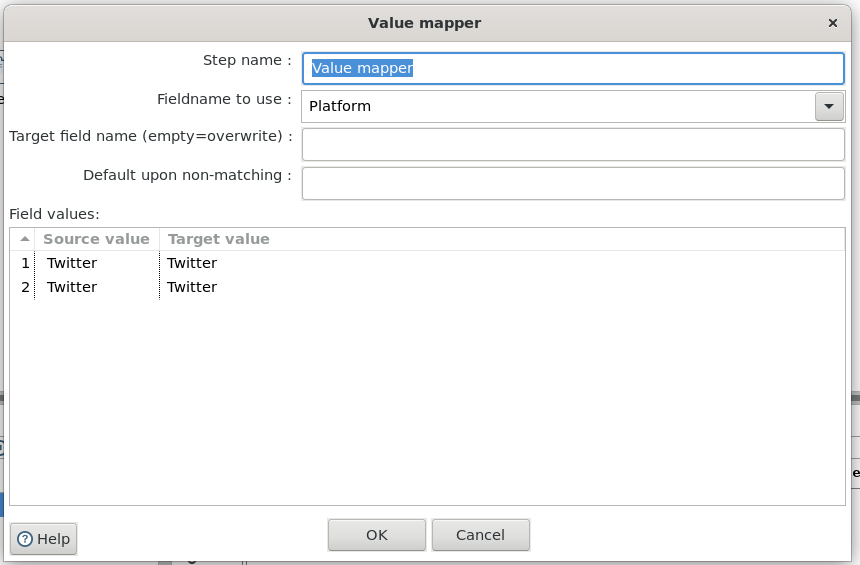
Потом надо изменить в поле Платформа значение слова Twitter. Так как в одном из слов в конце два пробела, а в другом один, надо сделать так, чтобы пробелов после Twitter не было, чтобы данные не дублировались (рисунок 5).

Рисунок 5. Изменение значений

После этого нужно подключиться к базе данных, выбрать таблицу Social\_Media\_Sentiments для дальнейшей выгрузки (рисунок 6).

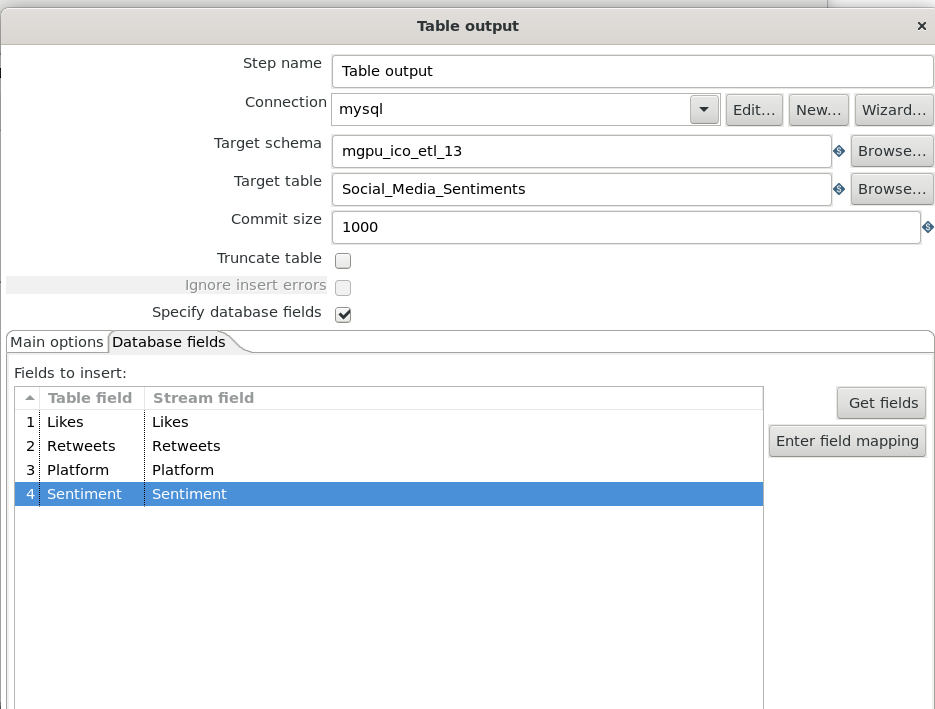


Рисунок 6. Подключение к базе данных

Все степы трансформации представлены на рисунке 7.

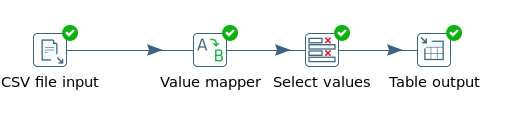


Рисунок 7. Степы трансформации

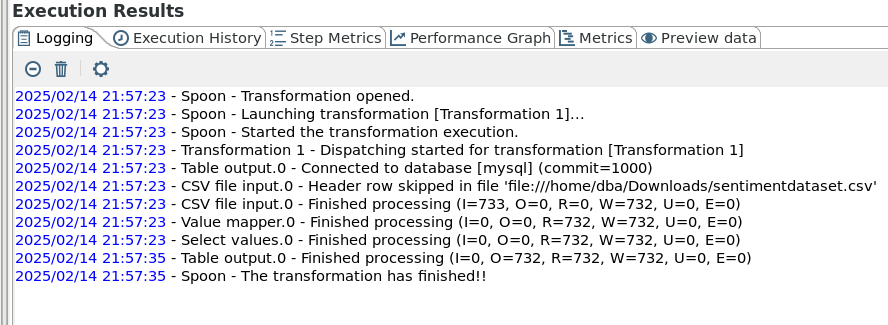
После запуска трансформации на рисунке 8 показана история логов. 

Рисунок 8. История логов

На рисунке 9 показаны метрики, а именно общее выполнение трансформации заняло 11678 миллисекунд: инициализация трансформации 152 миллисекунда, инициализация степа загрузка csv файла 0 миллисекунд, инициализация степа выбор значений 0 миллисекунд, инициализация степа изменение значений 0 миллисекунд, инициализация степа выгрузка в базу данных 135 миллисекунд, подключение к базе данных 121 миллисекунд, выполнение степа загрузка csv файла 34 миллисекунды, выполнение степа выбор значений 58 миллисекунд, выполнение степа изменение значений 52 миллисекунды, выполнение степа выгрузка в базу данных 11486 миллисекунд.

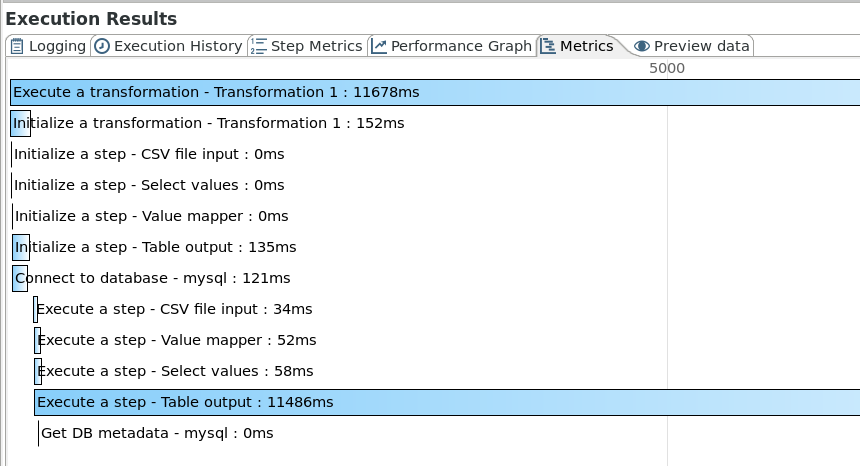


Рисунок 9. Метрики

На рисунке 10 показаны метрики степа.

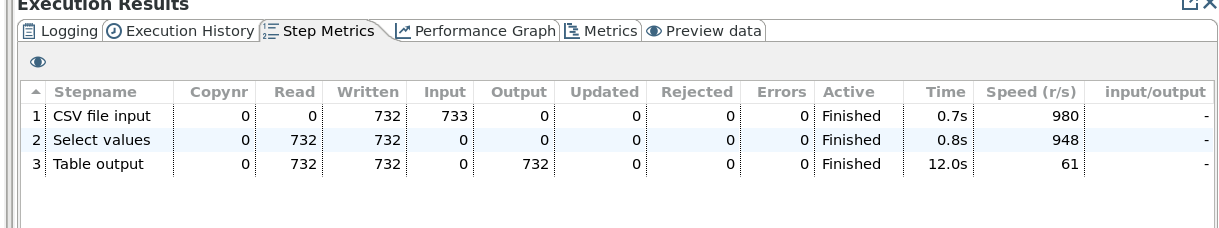


Рисунок 10. Метрики степа

График загрузки данных в Pentaho (рисунок 11).

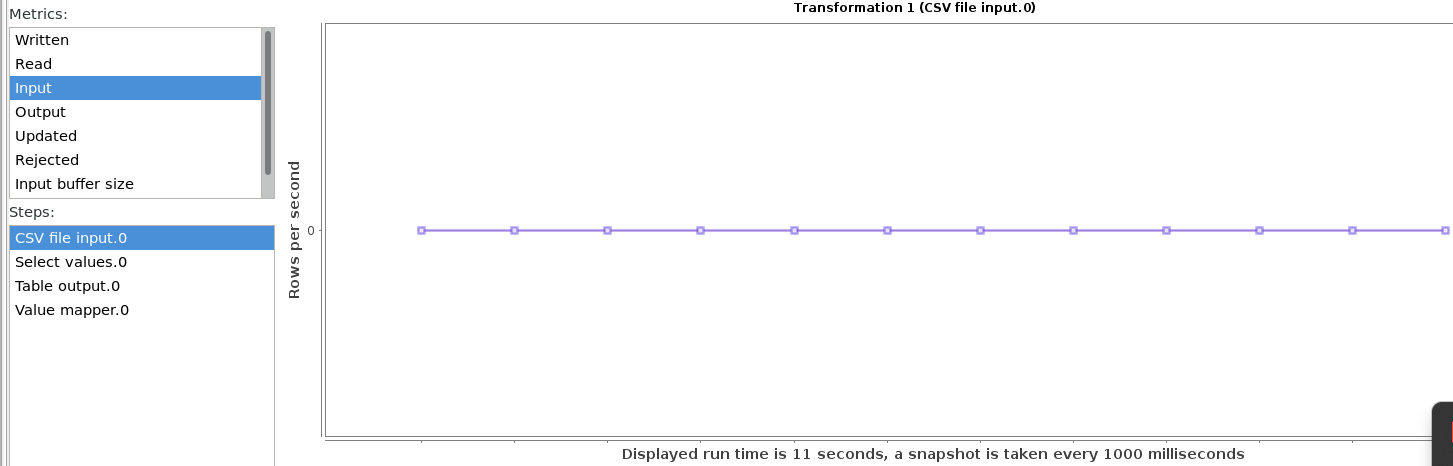


Рисунок 11. График загрузки данных в Pentaho

График выгрузки данных в базу данных (рисунок 12).

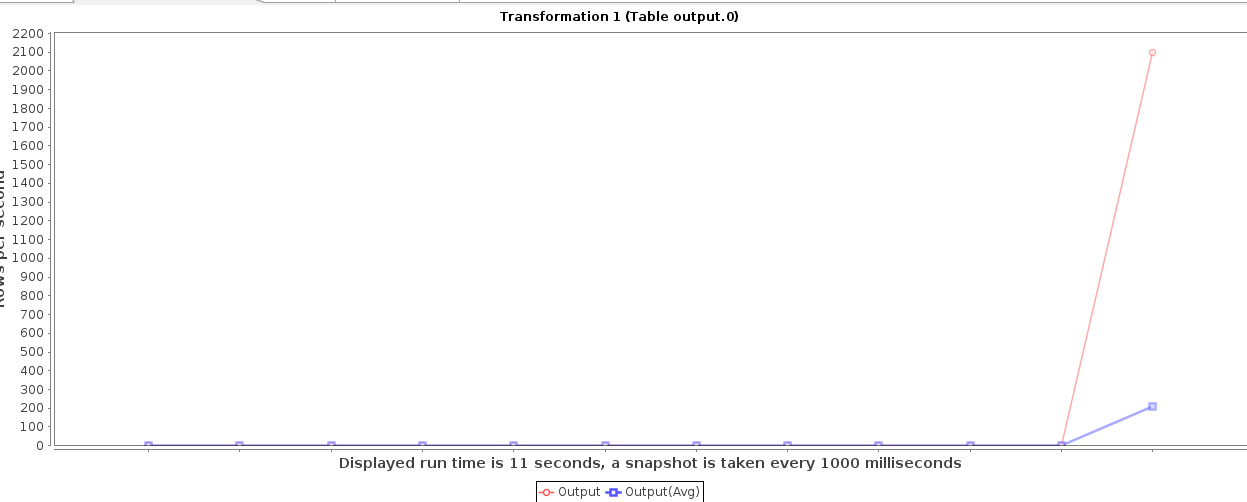


Рисунок 12. График выгрузки данных в базу данных

Как можно заметить на рисунке 13, все данные загрузились в PhpMyAdmin.

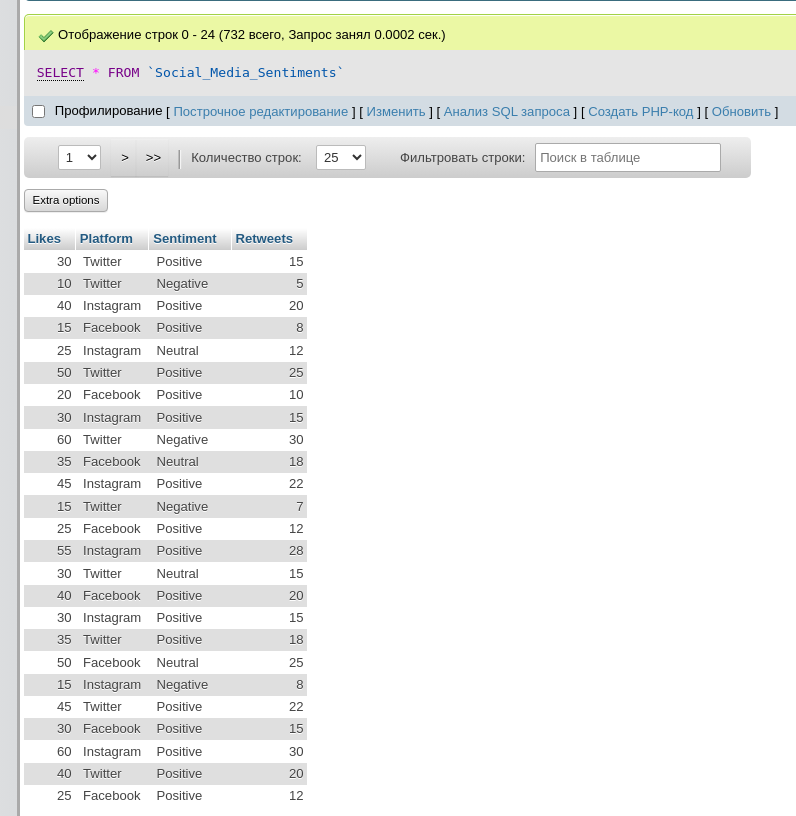


Рисунок 13. Данные успешно загрузились

Теперь надо выполнить SQL запросы. На рисунке 14 показано, сколько лайков на каждой платформе (запрос занял 0,0009 секунд). Больше всего в Instagram (соцсеть принадлежит компании Meta, признанной экстремистской в РФ), а меньше всего Facebook (соцсеть принадлежит компании Meta, признанной экстремистской в РФ).

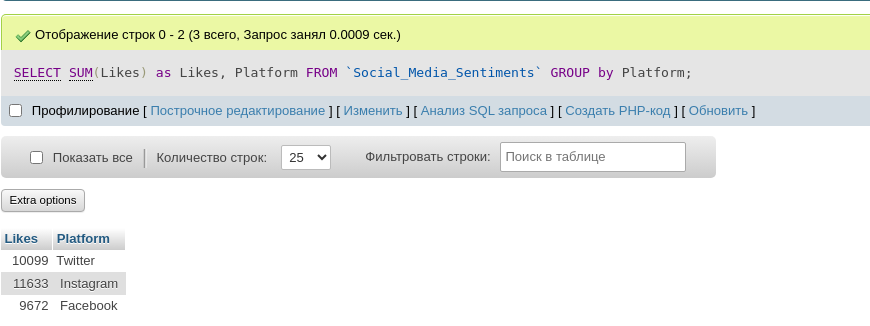


Рисунок 14. Количество лайков на каждой платформе

На рисунке 15 показано количество ретвитов на каждой платформе (запрос занял 0,0009 секунд).

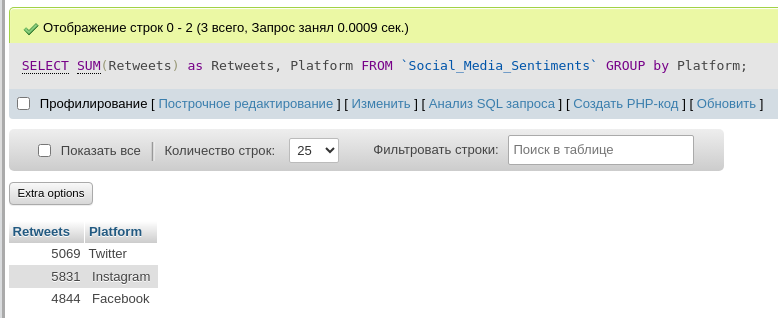


Рисунок 15. Количество ретвитов на каждой платформе

На рисунке 16 показано количество определенных реакций пользователей (запрос занял 0,0008 секунд). Больше всего позитивных реакций.

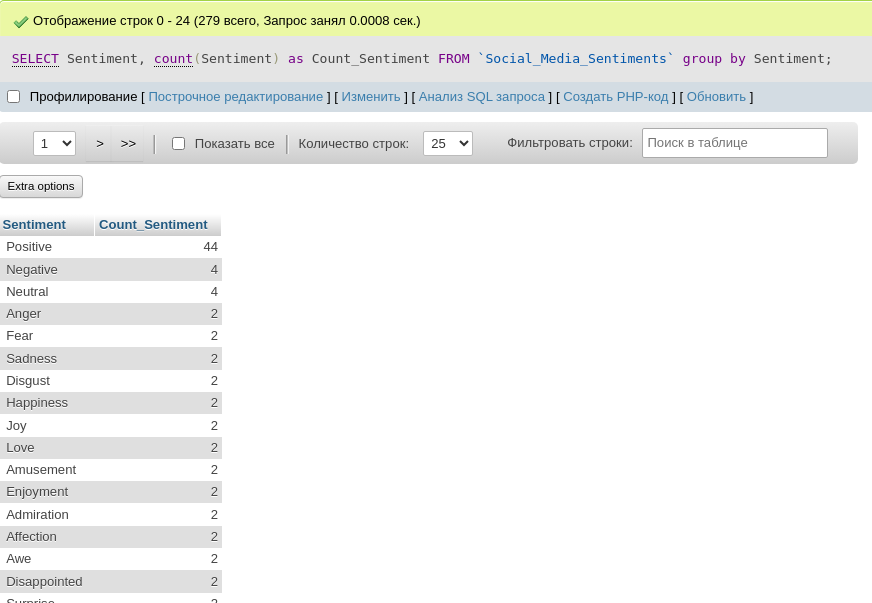


Рисунок 15. Количество определенных реакций пользователей

**Итоги работы:**

1. Был настроен и запущен Pentaho;
2. Было настроено подключение к базе данных;
3. Была реализована обработка данных активностей;
4. Были выполнены запросы в базе данных для проверки загруженных данных.