**Цель работы:** получить практические навыки создания ETL-процесса для загрузки данных из CSV-файла в базу данных MySQL с использованием Pentaho Data Integration.

**Задачи:**

− Создать динамические подключения к различным источникам данных.

− Разработать процесс выявления и обработки дублирующихся записей.

− Реализовать механизм объединения данных в единое хранилище.

− Настроить обработку ошибок при выполнении трансформации. Программное обеспечение: - Pentaho Data Integration 9.4. - MySQL или PostgreSQL. - CSV или Excel файлы с тестовыми данными.

**Вариант 13. Фильтр по штату: только Texas, анализ возвратов, отчет по доставке**

Job показан на рисунке 1.

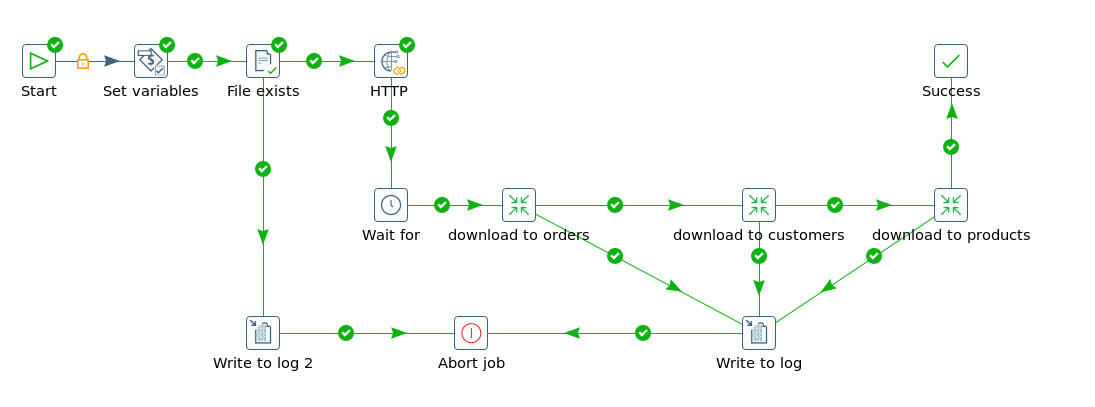
****

Рисунок 1. Job

Поток orders продемонстрирован на рисунке 2.

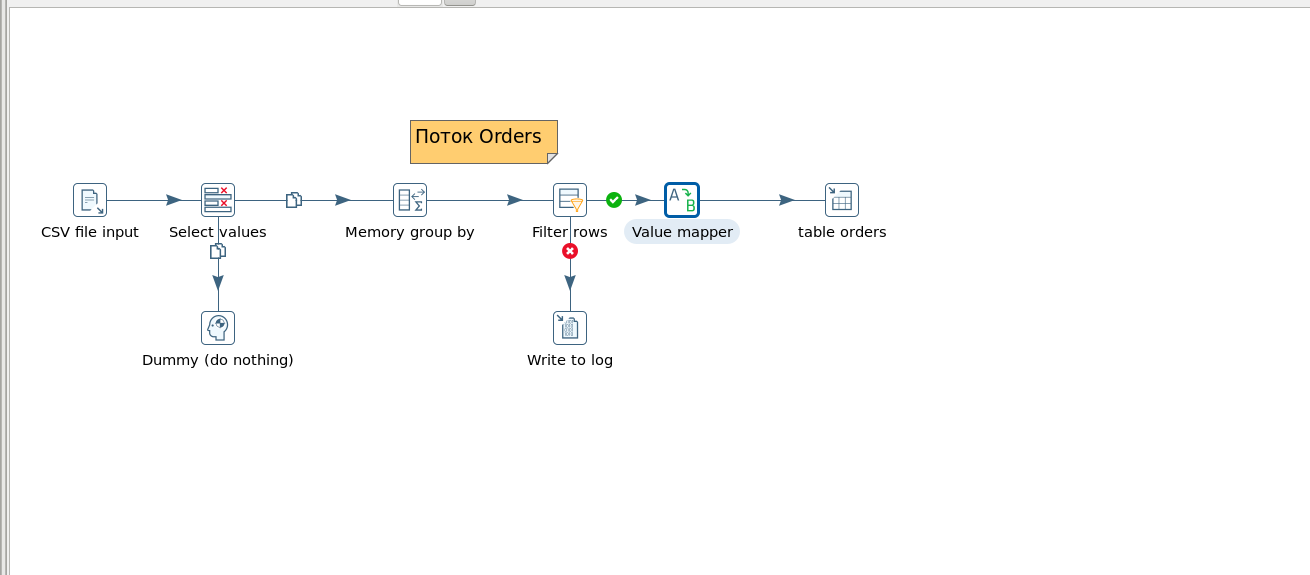
****

Рисунок 2. Поток Orders

На рисунке 3 показана трансформация Customers

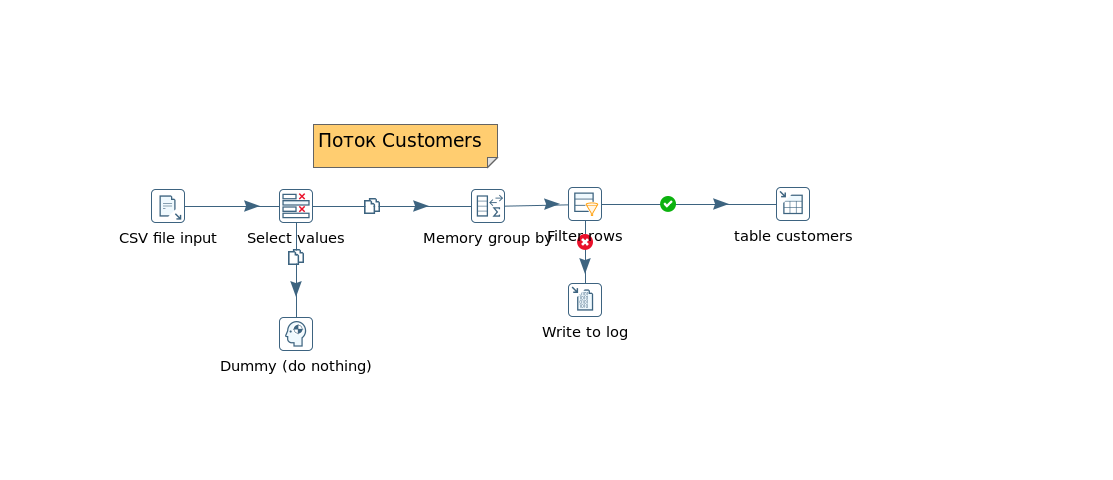
****

Рисунок 3. Трансформация Customers

На рисунке 4 показан фильтр, где нужен Техас

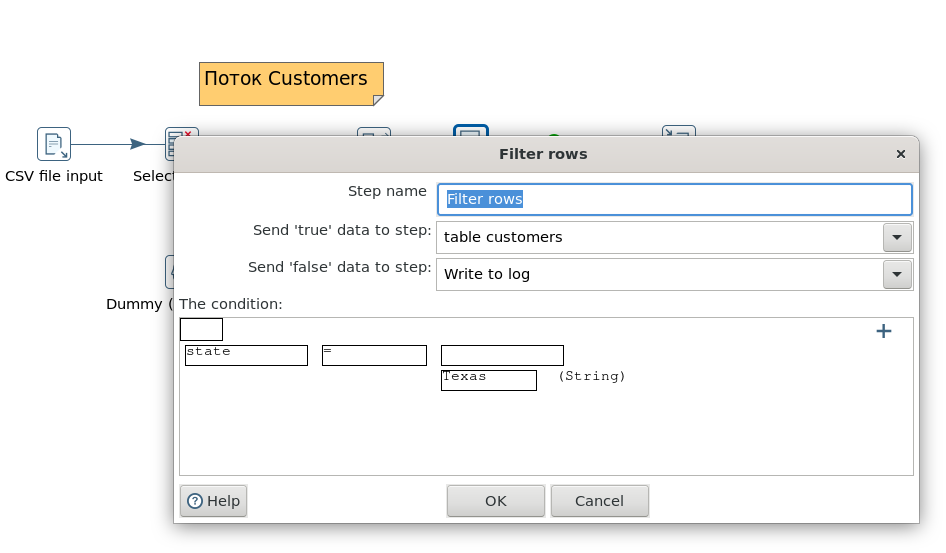
****

Рисунок 4. Фильтр Техас

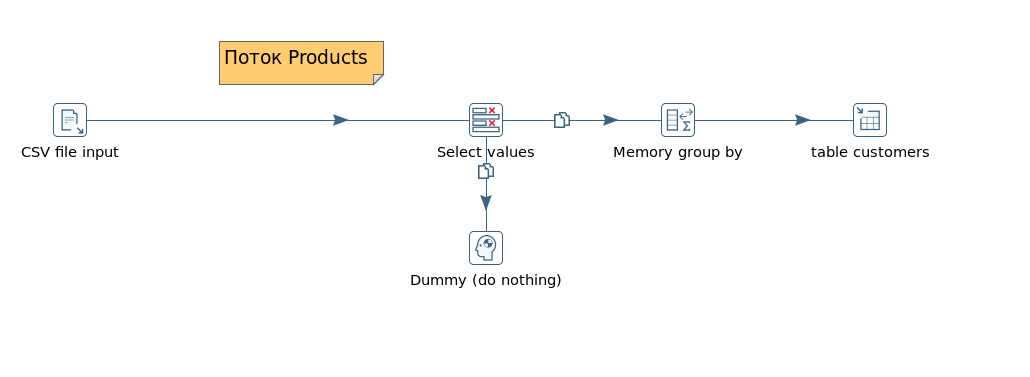
На рисунке 5 показана трансформация Products.****

Рисунок 5. Трансформация Products

На рисунках 6-8 продемонстрировано, что все успешно загрузилось в базу данных

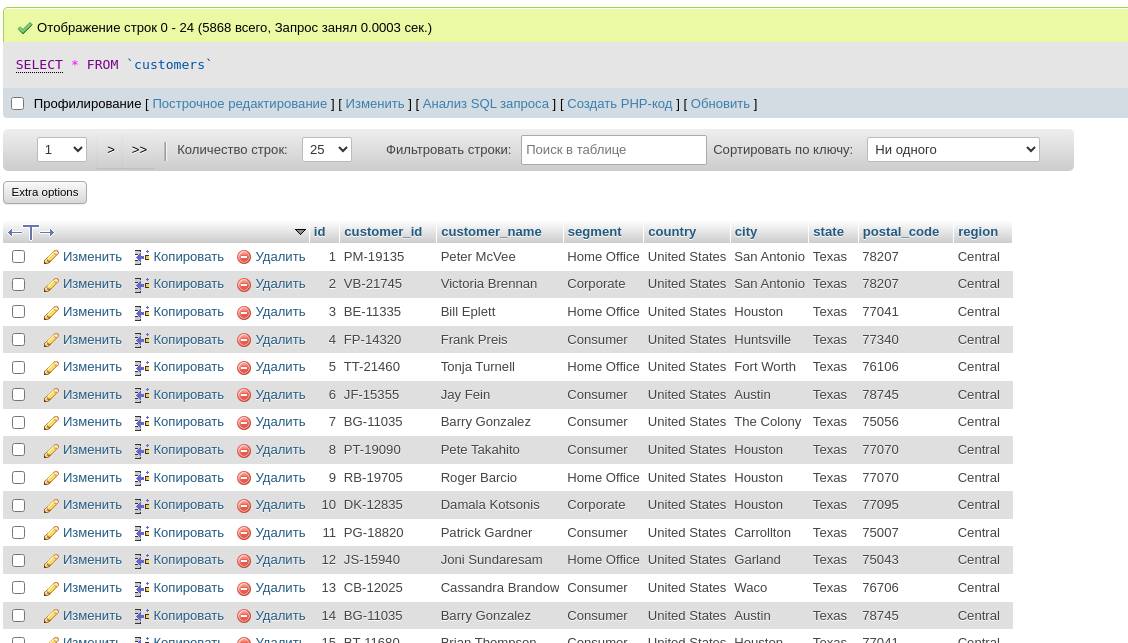
****

Рисунок 6. Загрузка в базу данных таблицы Customers

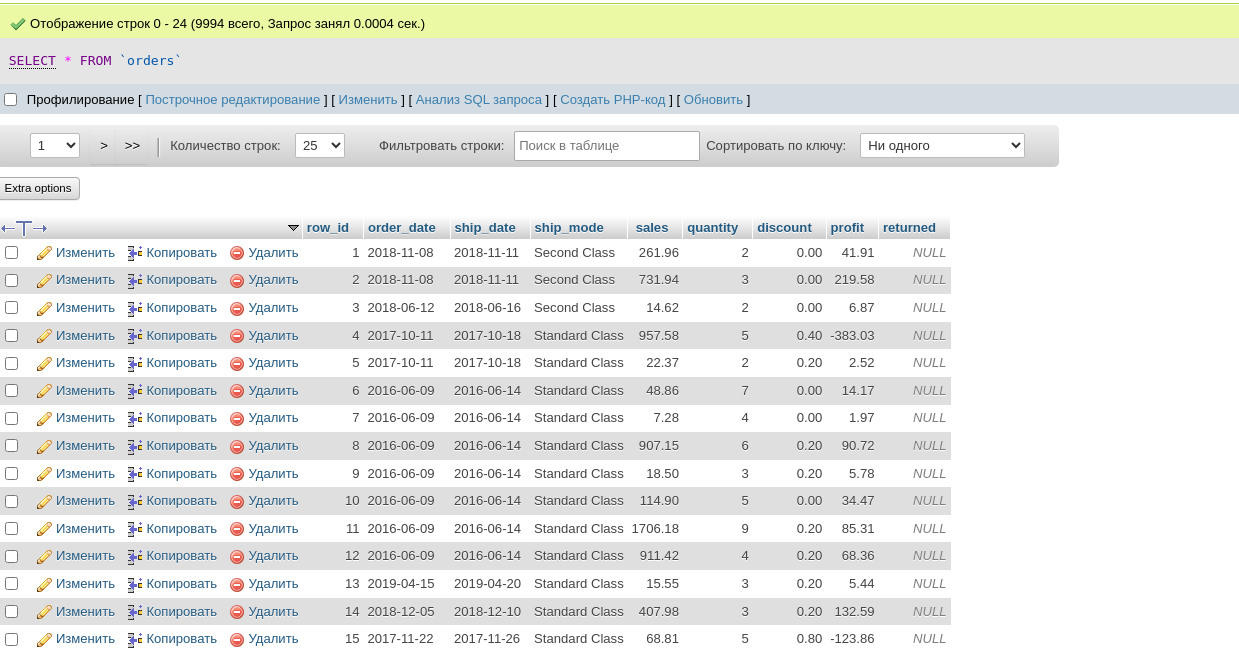
****

Рисунок 7. Загрузка в базу данных таблицы Orders

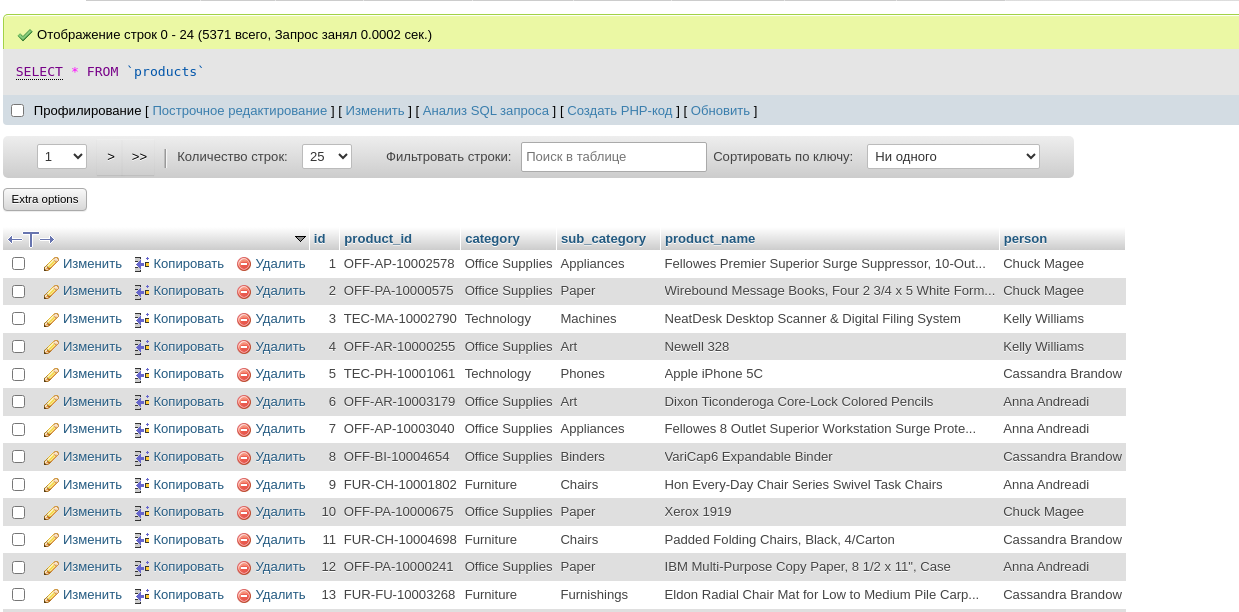
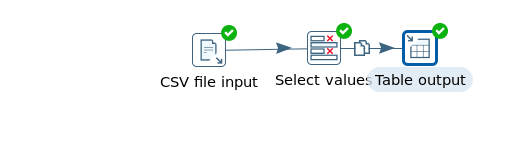
****

Рисунок 8. Загрузка в базу данных таблицы Products

Затем нужно сделать анализ по возвратам. Трансформация представлена на рисунке 9.

  
Рисунок 9. Трансформация по возвратам

На рисунке 10 показана выгрузка в базу данных

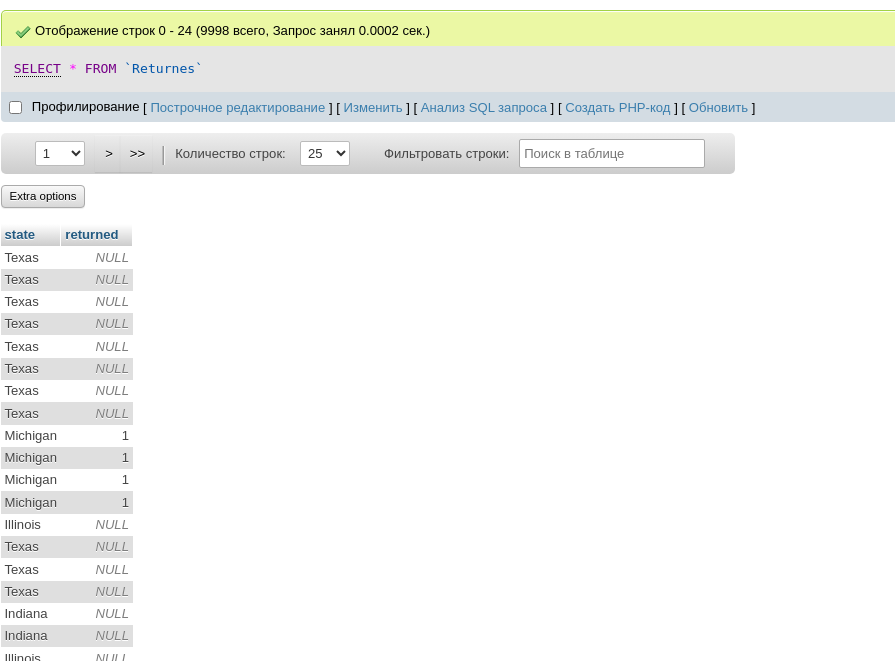


Рисунок 11. Выгрузка в базу данных

На рисунке 12 показан анализ по возвратам. Больше всего возвратов у штата Калифорния.

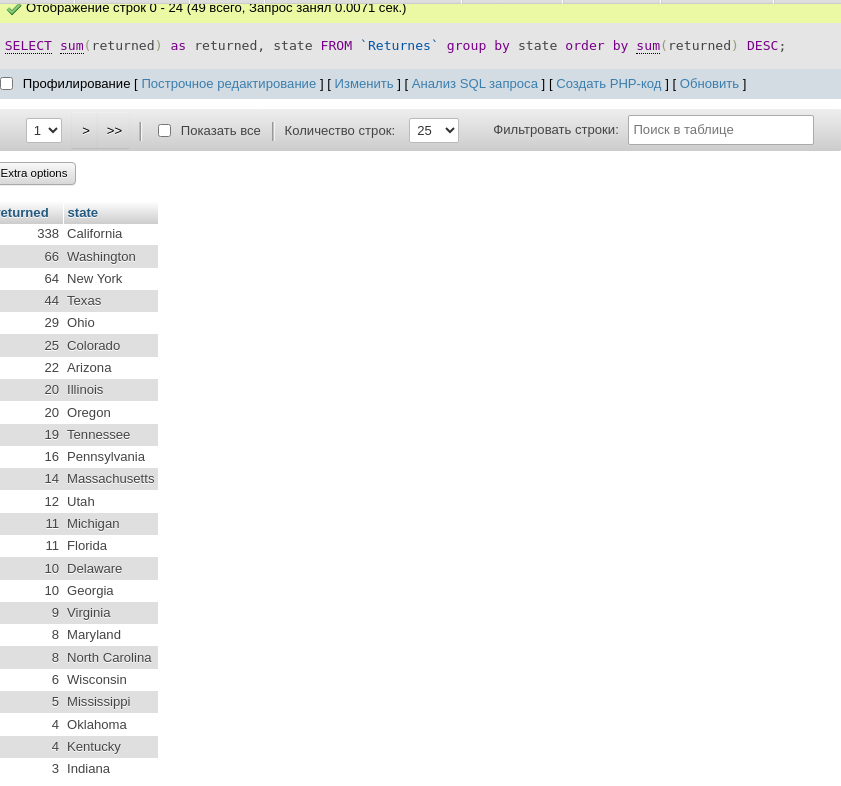


Рисунок 12. Анализ возвратов у разных штатов

Далее идет анализ по доставкам. Трансформация представлена на рисунке 13.

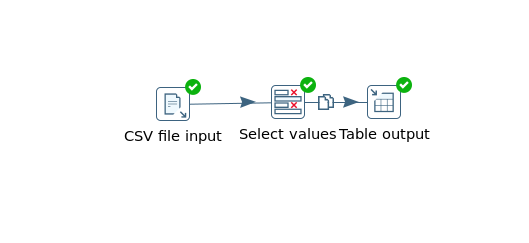


Рисунок 13. Трансформация по доставкам

Как можно заметить на рисунке 14, все выгрузилось

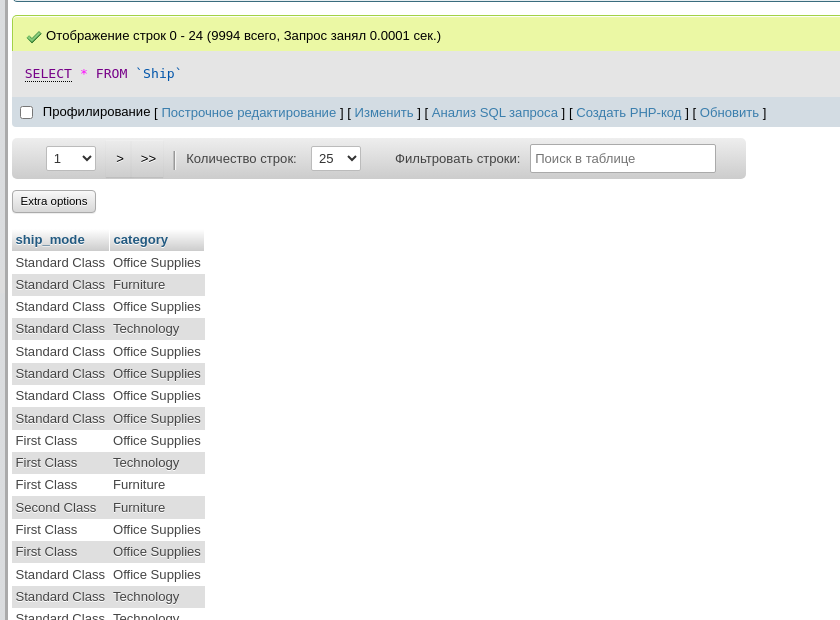


Рисунок 14. Загрузка в базу данных

На рисунке 15 можно увидеть количество доставок. Больше всего стандартного класса

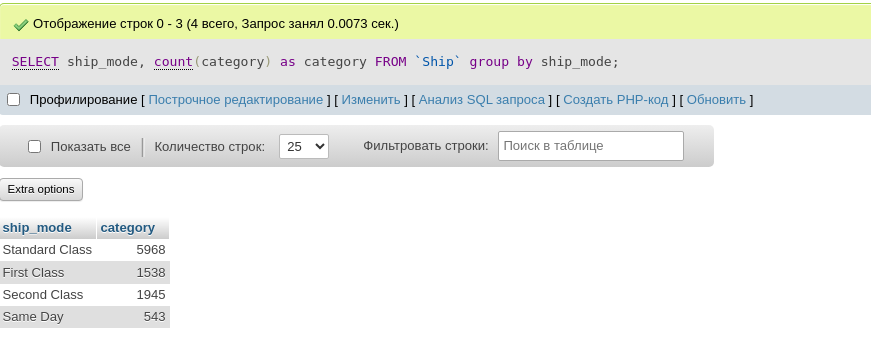


Рисунок 15. Количество доставок

**Выводы:**

1. Динамические подключения к различным источникам данных были реализованы
2. Разработан процесс выявления и обработки дублирующихся записей.
3. Реализован механизм объединения данных в единое хранилище