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 №2-1

Тема:

«Динамические соединения с базами данных»

Выполнил(а): Морозова Валерия АДЭУ-211

Преподаватель:

Москва

2025

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.

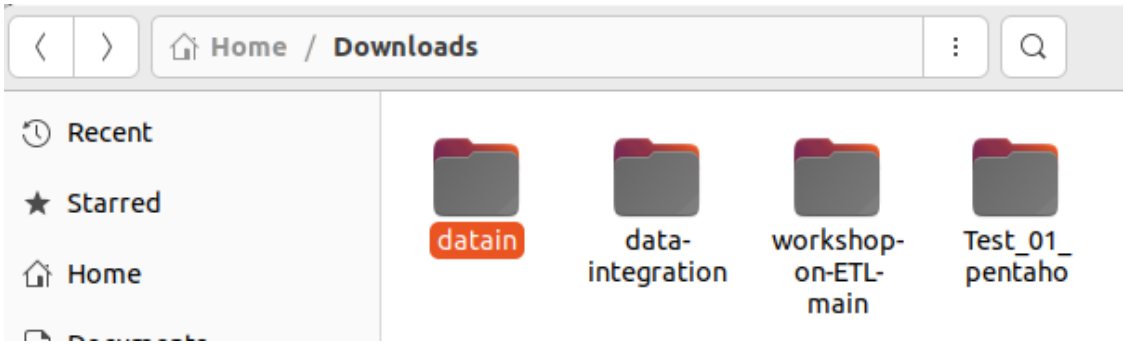

Рисунок 1. Создание директории datain

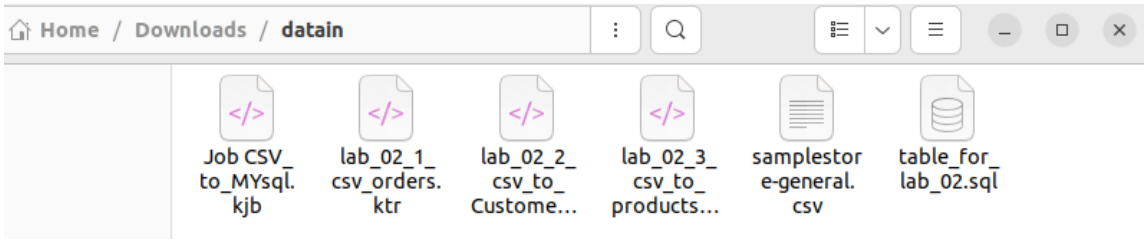

Рисунок 2. Пополнение директория необходимыми файлами

Теперь необходимо открыть их и запустить в Pentaho предварительно изменив пути

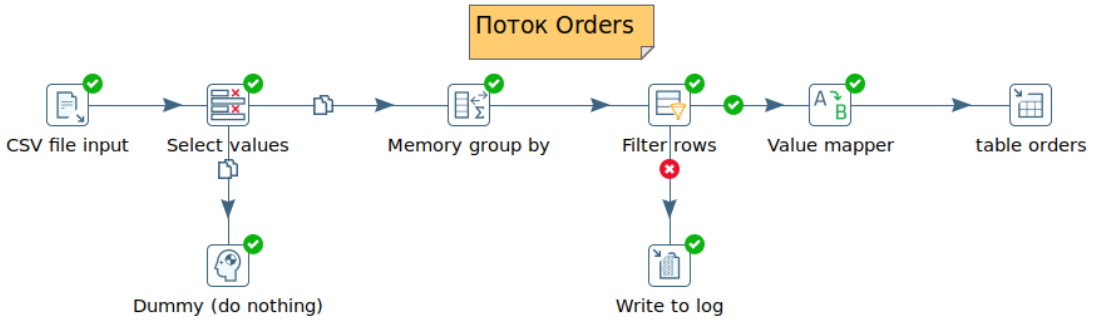

Рисунок 3. Трансформация lab_02_1_csv_orders.ktr

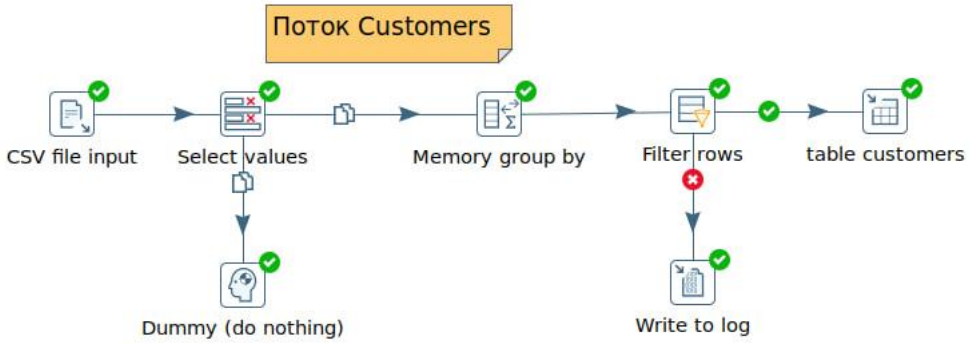

Рисунок 4. Трансформация lab_02_2_csv_customers.ktr

Трансформация с добавленным фильтром по стране: только United States (Вариант 10).

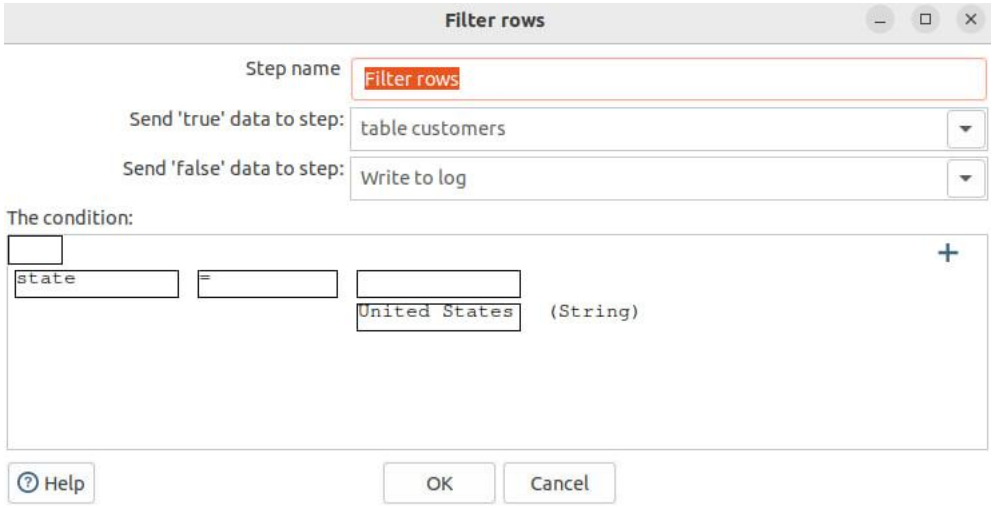

Рисунок 5. Фильтр только United States

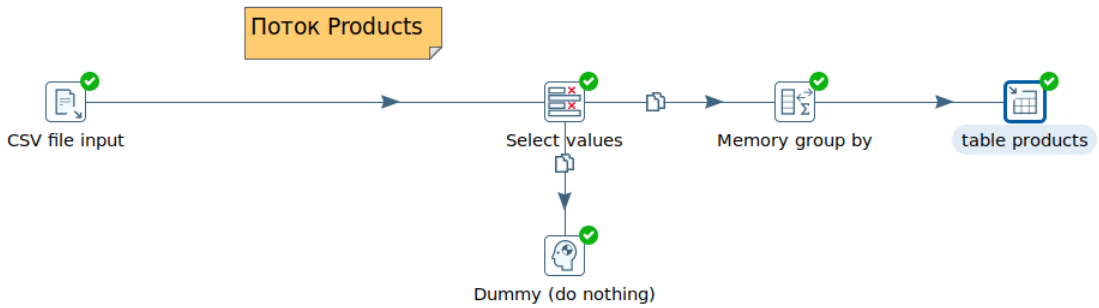

Рисунок 6. Трансформация lab_02_3_csv_products.ktr

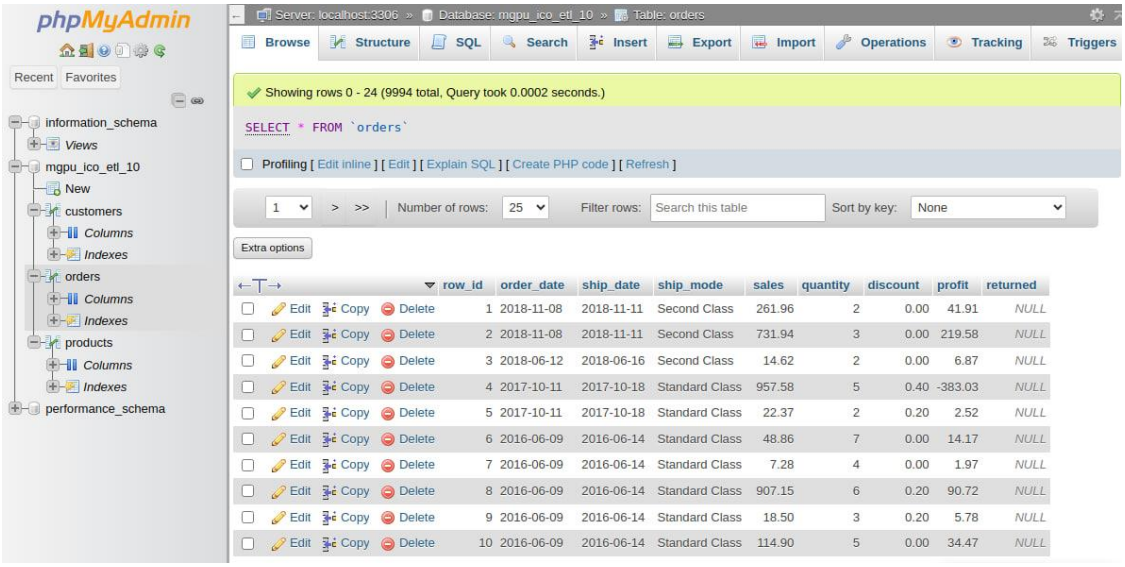

Рисунок 7. В MySQL была создана и заполнена данными таблица orders

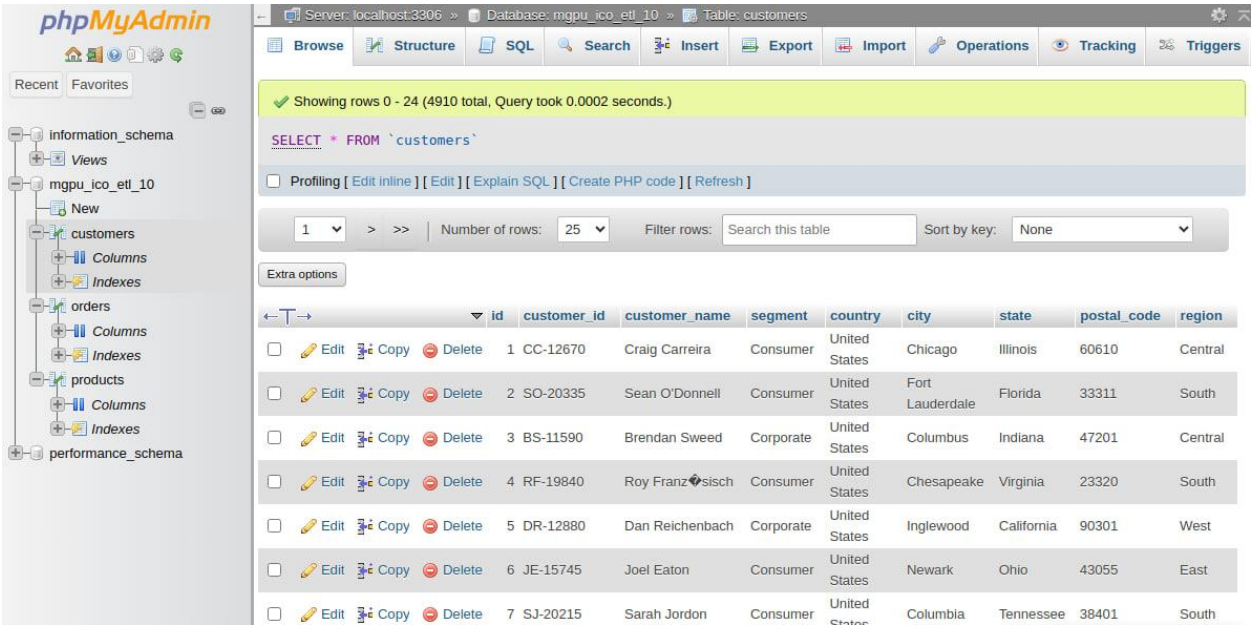

Рисунок 8. В MySQL была создана и заполнена данными таблица customers

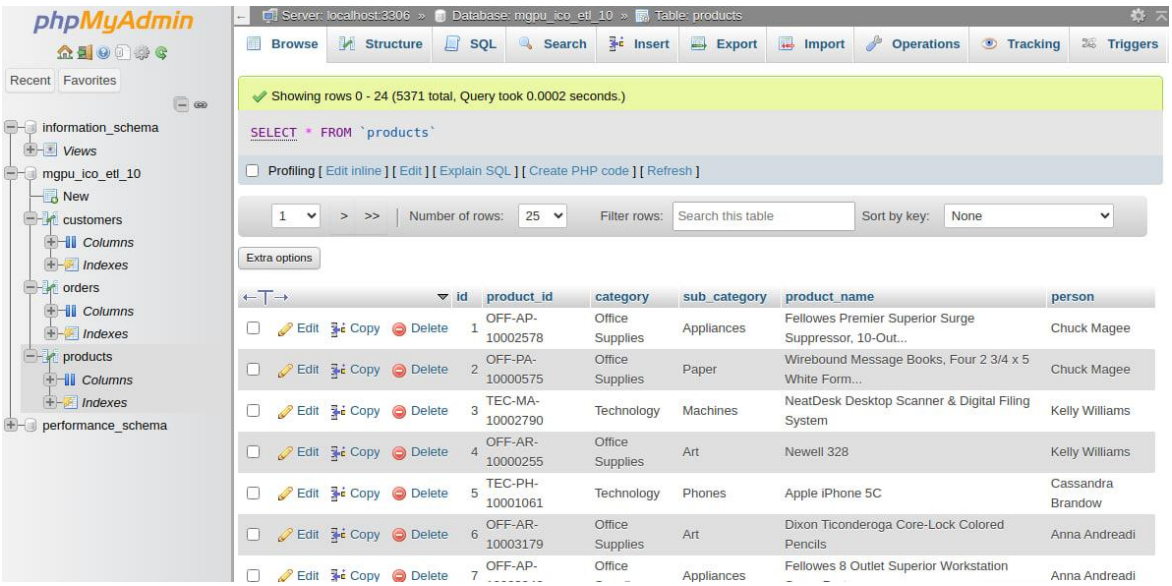

Рисунок 9. В MySQL была создана и заполнена данными таблица products

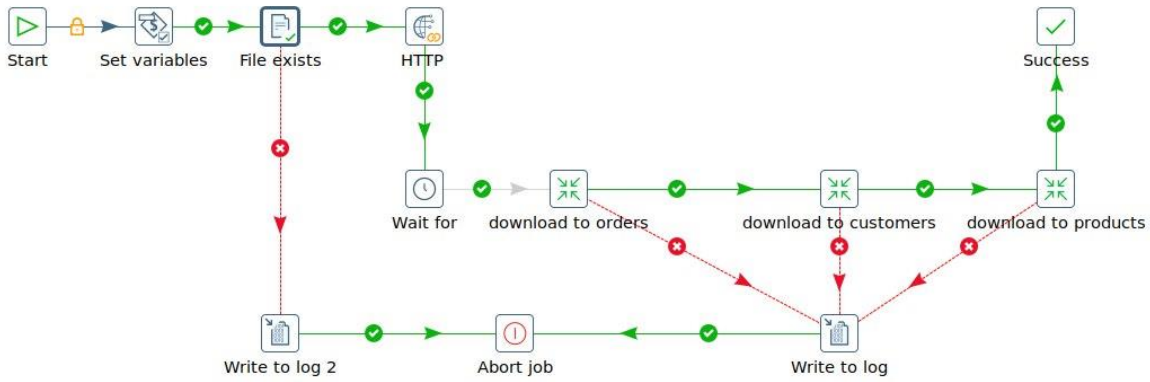

Рисунок 10. Job CSV_to_MYsql отработал после загрузки файла по протоколу HTTP

Индивидуальные задания

Анализ скидок

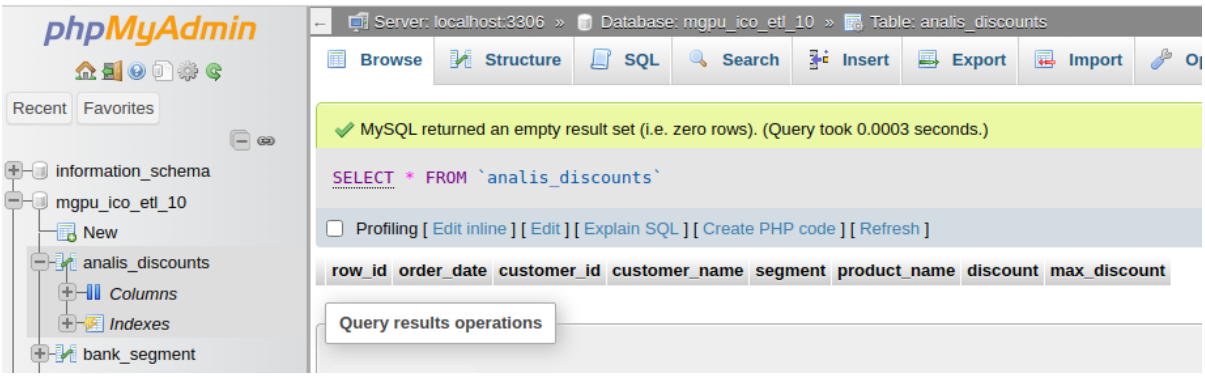

Рисунок 11. Создала таблицу с нужными столбцами для анализа скидок

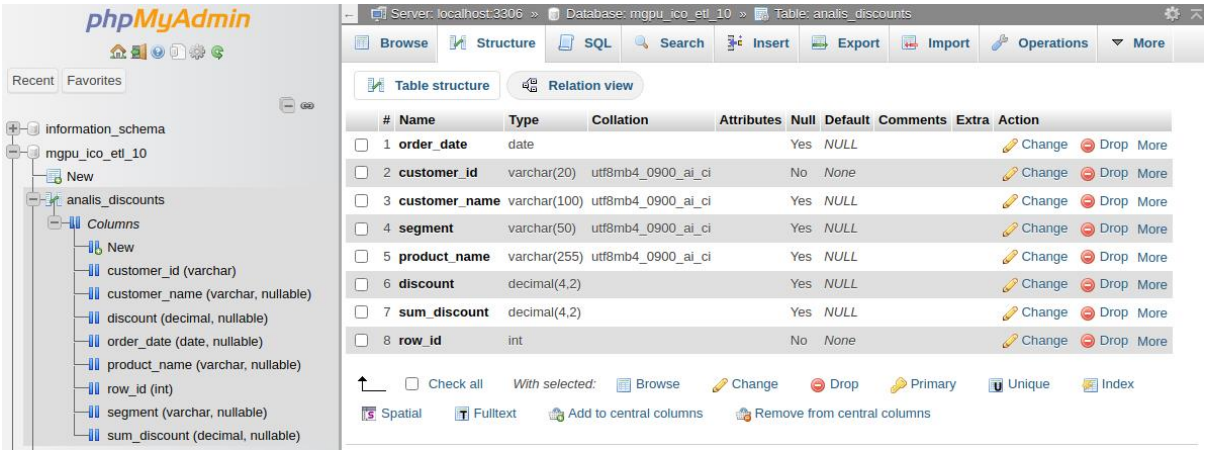

Рисунок 12. Скорректировала типы данных

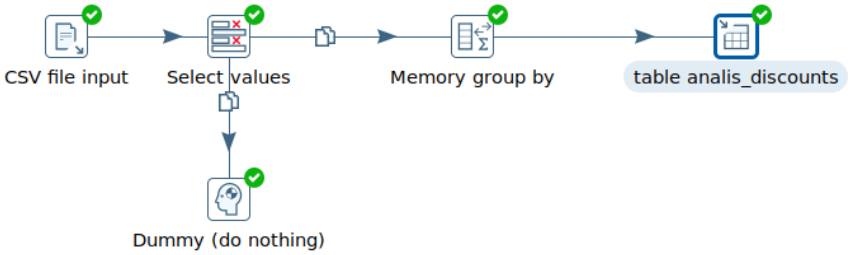

Рисунок 13. Трансформация lab_02_4

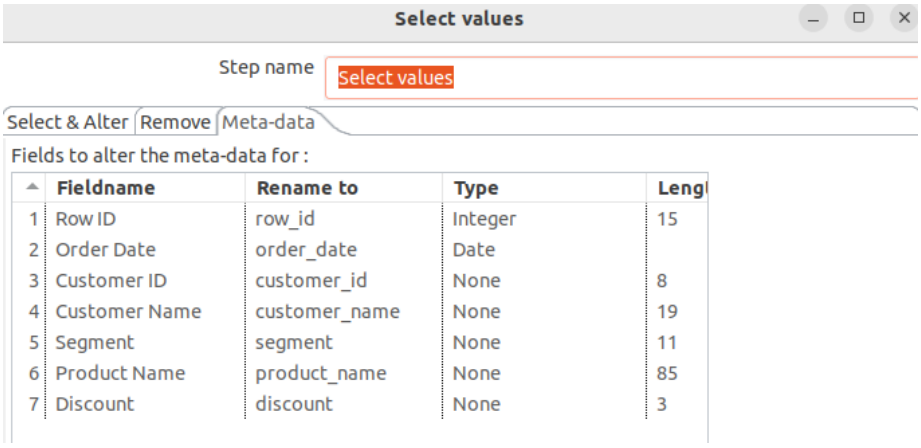

Рисунок 14. Выбор нужных столбцов

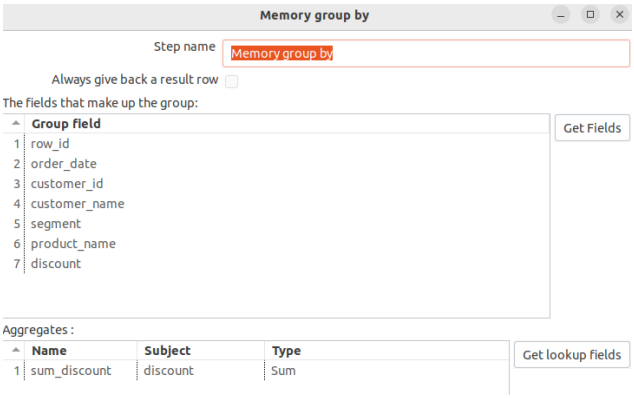

Рисунок 15. Группировка по сумме скидок

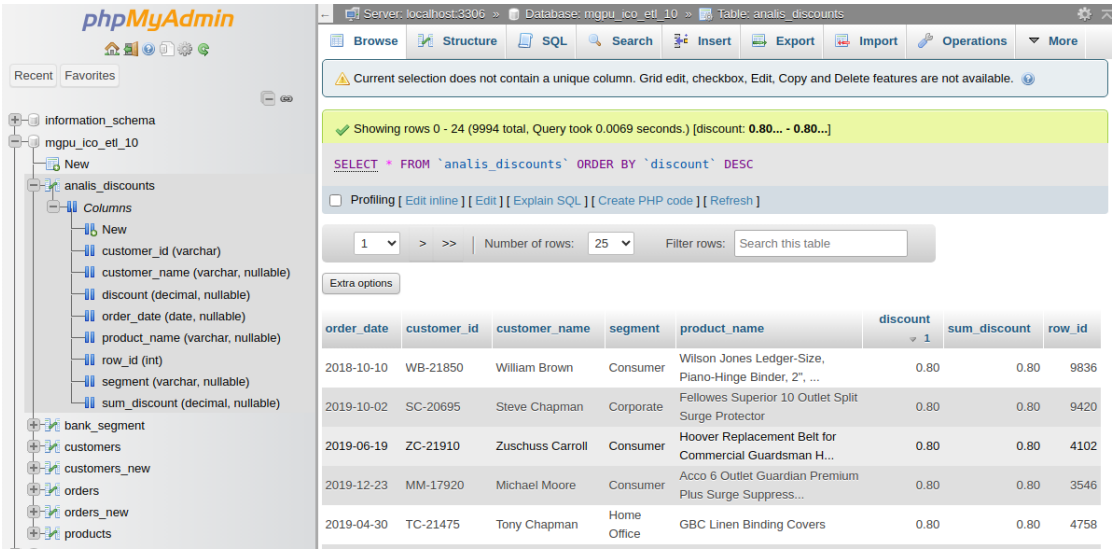

Рисунок 16. Данные загружены в таблицу analis_discount

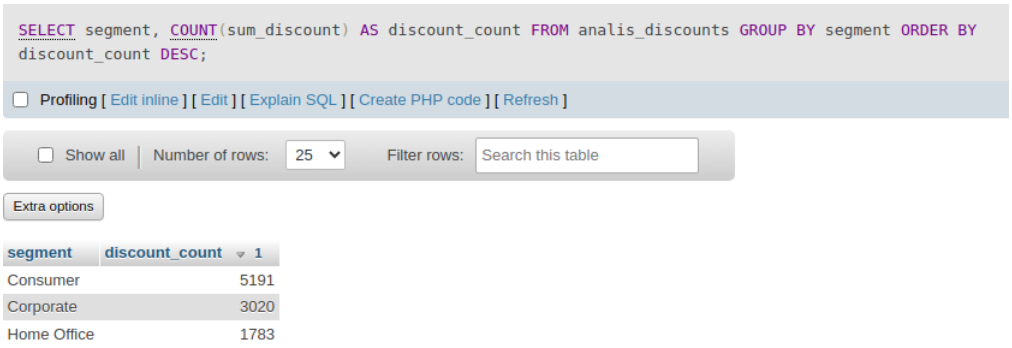

Рисунок 17. Количество скидок для разных сегментов потребителей

Как видно, на первом месте по количеству получаемых скидок и акций на товары у обычных потребителей, на втором месте корпорации и на последнем домашний офис.

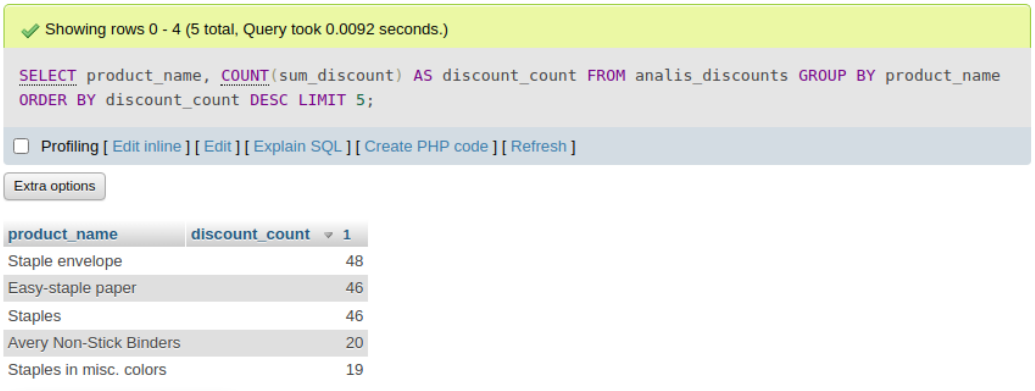

Рисунок 18. На какие товары чаще всего устанавливают скидки

Отчет по регионам

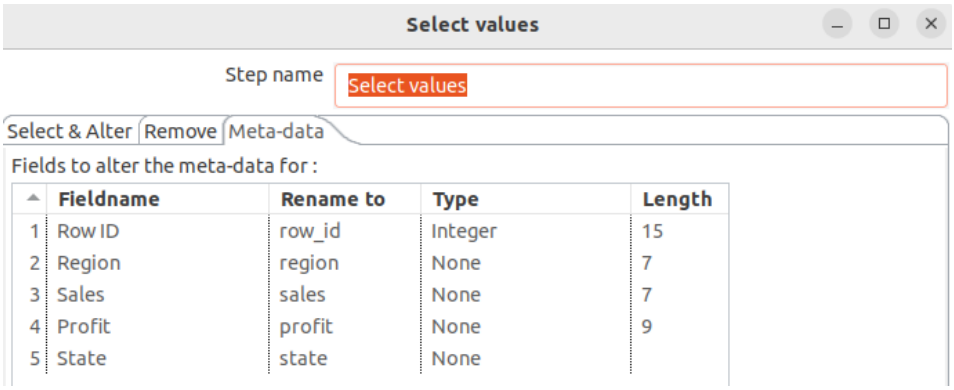

Рисунок 19. Выбор нужных столбцов для анализа

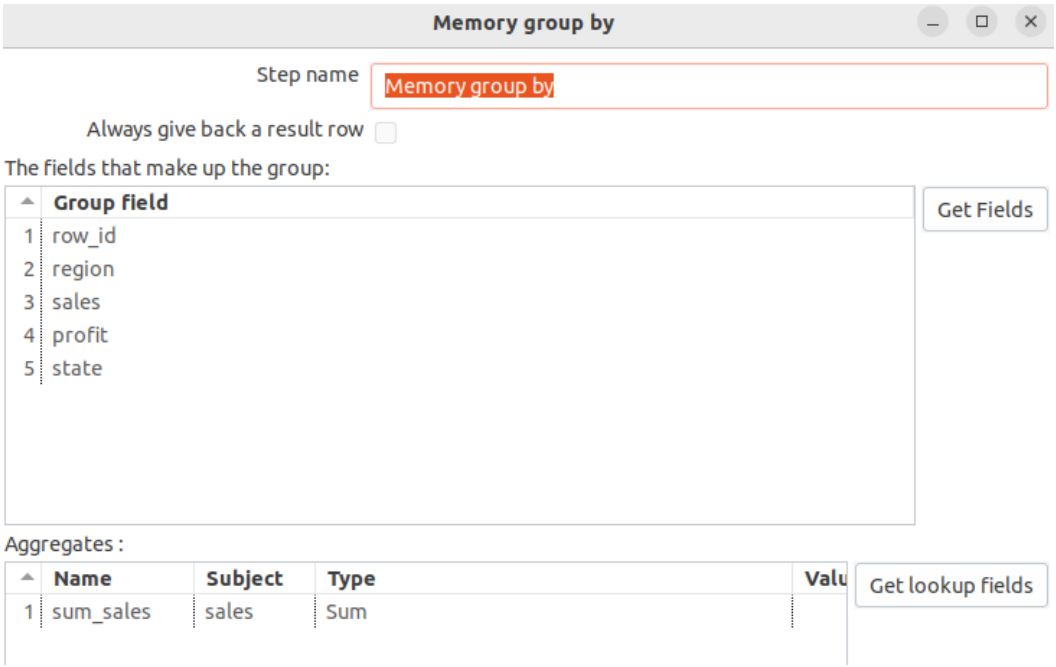

Рисунок 20. Группировка значений по сумме продаж

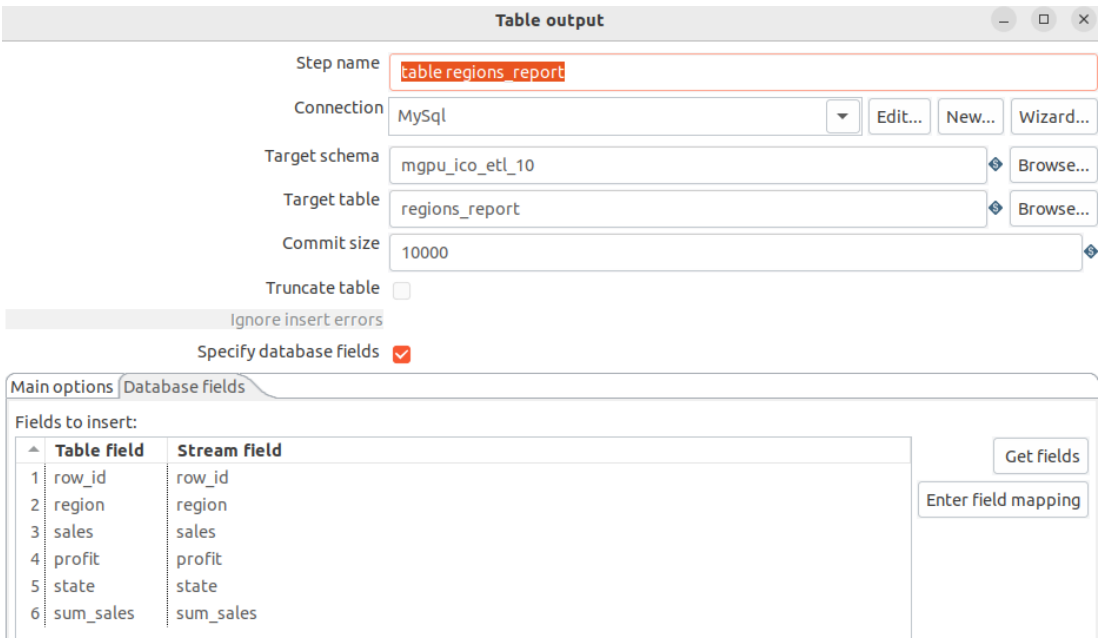

Рисунок 21. Выгрузка данных в таблицу regions_report

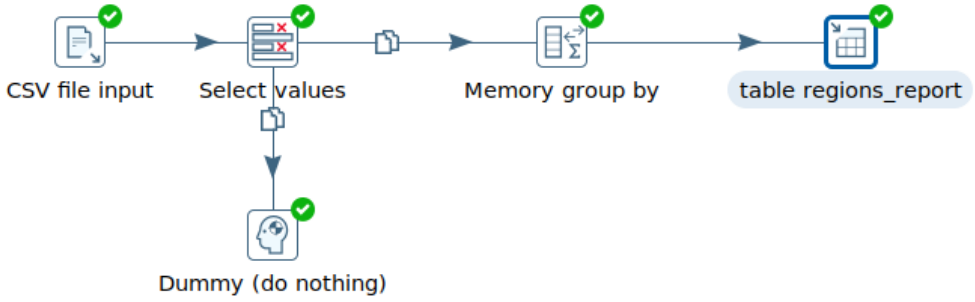

Рисунок 22. Трансформация lab_02_5

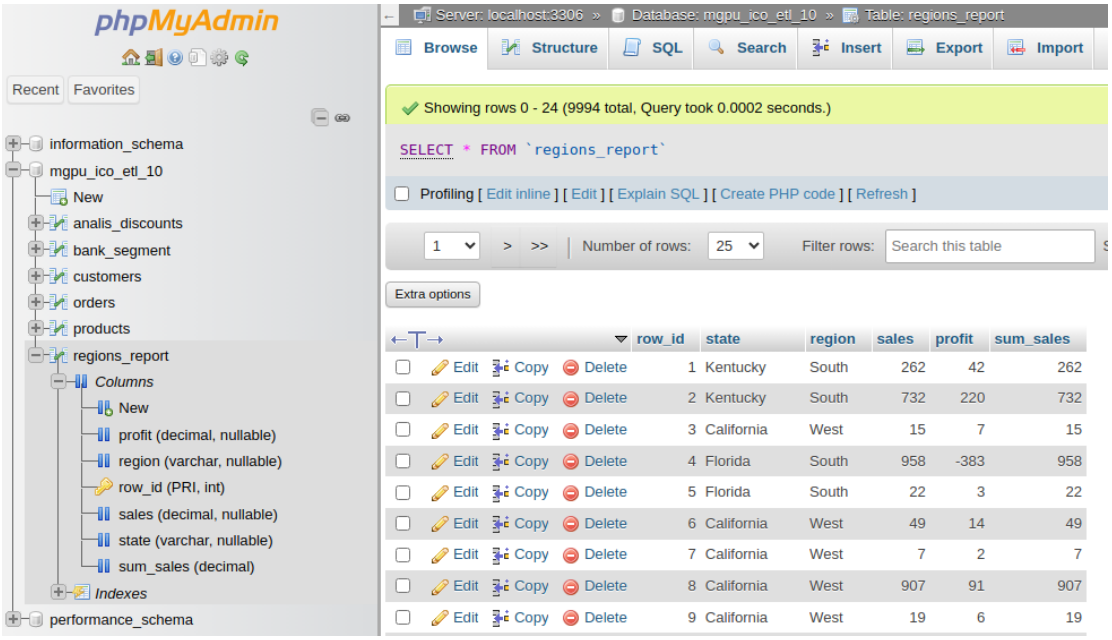

Рисунок 23. Данные загружены в таблицу regions_report

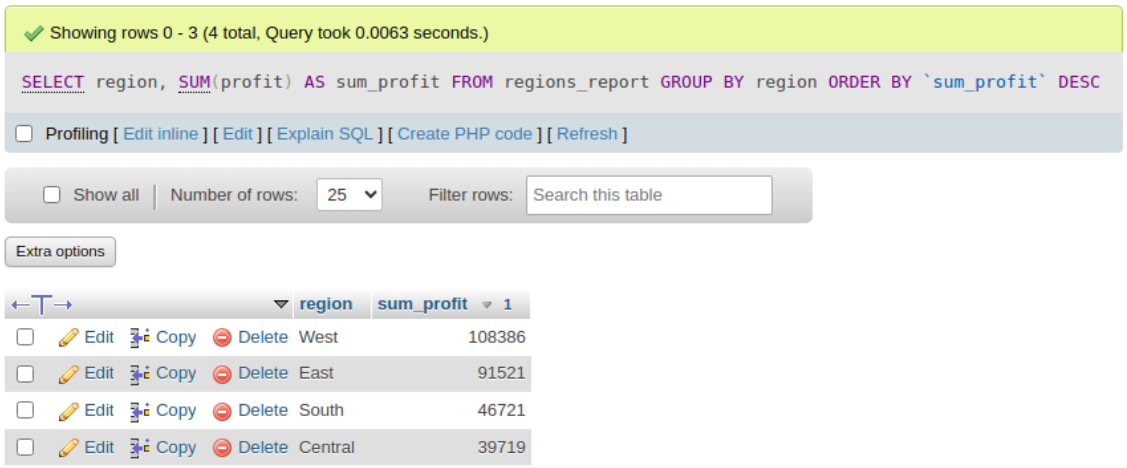

Рисунок 24. Прибыль по регионам

Согласно sql запросу видно, что больше всего выручки делает западный регион, а меньше всего центральный.

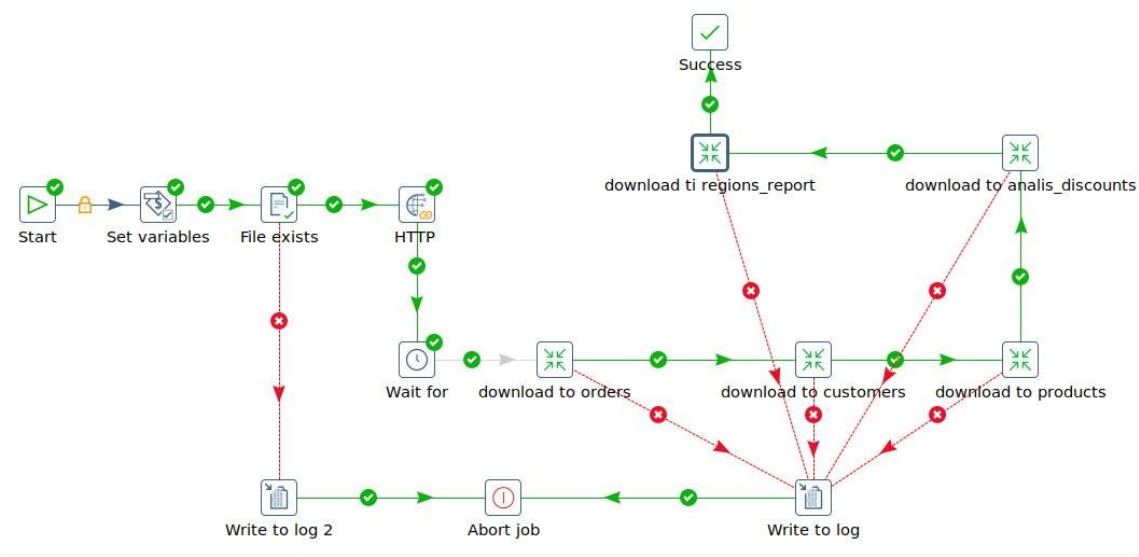

Рисунок 25. Итоговый Job CSV_to_MYsql с двумя новыми трансформациями

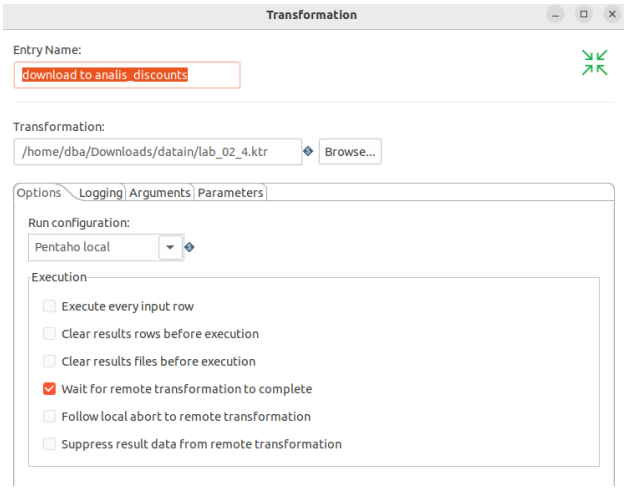

Рисунок 26. Настройка степа трансформации по анализу скидок

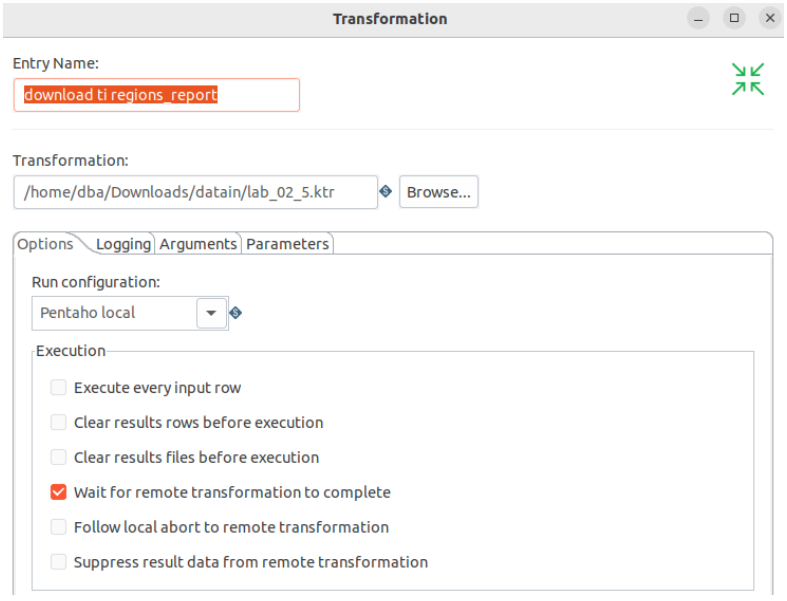

Рисунок 26. Настройка степа трансформации по отчету по регионам

Выводы

Выполнены все трансформации

Успешно выполнена и дополнена двумя трансформациями, согласно индивидуальному заданию, структура job:

Start

├── Set Variables (FILE_PATH)

├── Check File Exists

├── HTTP Download

├── Transform: Load Orders (lab_02_1)

├── Transform: Load Customers (lab_02_2)

├── Transform: Load Products (lab_02_3)

├── Transform: Analytics 1 (lab_02_4)

└── Transform: Analytics 2 (lab_02_5)

Проведен анализ скидок, согласно которому, видно, что более всего потребителями и корпорациями востребованы срепочные конверты, легко скрепляемая бумага, скобы и антипригарные папки, соответсвенно больше всего скидок именно на эти товары.

Сделан отчет по регионам, где видно, что больше всего выручки делает западный регион, а меньше всего центральный.

Контрольные вопросы

1. Что такое динамические соединения в Pentaho Data Integration?

Динамические соединения в Pentaho Data Integration позволяют создавать подключения к базам данных или другим источникам данных на лету, основываясь на значениях, полученных во время выполнения трансформации. Это особенно полезно, когда нужно подключаться к различным источникам данных в зависимости от условий, таких как параметры, заданные пользователем, или результаты предыдущих шагов трансформации. Используя динамические соединения, можно избежать необходимости жестко задавать параметры подключения в трансформации.

2. Как организовать обработку ошибок в трансформации?

Обработка ошибок в PDI может быть организована с помощью нескольких методов:

"Error Handling": Многие компоненты (например, "Table Output" или "Text File Output") имеют встроенные опции для обработки ошибок, такие как запись ошибок в отдельный файл или таблицу.

"Error Rows": Вы можете настроить поток данных так, чтобы ошибки направлялись в отдельный поток, например, через "Filter Rows" или "Switch/Case", где можно обрабатывать их отдельно.

"Try-Catch": В PDI также доступны "Try-Catch" блоки, которые позволяют обрабатывать исключения и выполнять альтернативные действия в случае ошибок.

3. Какие методы выявления дублей существуют?

Существует несколько методов выявления дублей в PDI:

"Unique Rows": Этот шаг позволяет удалить дубликаты на основе заданных полей.

"Group By": С помощью этого шага можно агрегировать данные и выявить дубликаты, используя группировку по определённым полям.

"Row Normalizer": Этот шаг позволяет нормализовать строки и выявить дублирующиеся записи.

Сравнение данных: Можно использовать "Merge Join" или "Join" для сравнения данных из разных источников и выявления дублей.

4. Как настроить параметризацию подключений?

Параметризацию подключений в PDI можно настроить следующим образом:

- ❖ Создание параметров: В диалоговом окне "Transformation" или "Job" можно создать переменные, которые будут использоваться в качестве параметров.
- ❖ Использование переменных: В настройках подключения к базе данных указать параметры подключения (например, имя пользователя,

пароль, URL) с использованием переменных, например, `${db_user}` или `${db_password}`.

❖ Передача значений: При запуске трансформации или задания можно передавать значения для этих переменных, что позволит динамически изменять параметры подключения.

5. Какие компоненты Pentaho Data Integration используются для объединения данных?

Для объединения данных в PDI используются следующие компоненты:

"Merge Join": Позволяет объединять данные из двух потоков на основе общего поля.

"Join Rows": Этот шаг также позволяет объединять строки из разных источников, используя заданные ключи.

"Union": Объединяет строки из нескольких источников в один поток.

"Group By": Хотя в первую очередь используется для агрегации, этот компонент также может помочь в объединении данных, группируя их по определённым полям.