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 №1-1**

**Тема:**

## «Установка и настройка ETL-инструмента. Создание конвейеров данных»

Выполнил(а): Морозова Валерия АДЭУ-211

Преподаватель:

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 Excel файле, 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.

Развертывание Pentaho Data 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.

**Вариант 10.** Анализ банковских транзакций: выявление паттернов, сегментация

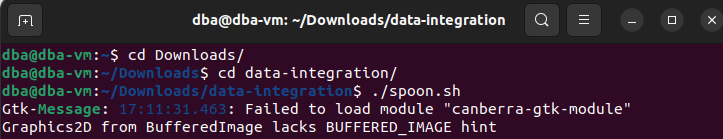


Рисунок 1. Переход в нужную папку и запуск Pentaho Spoon

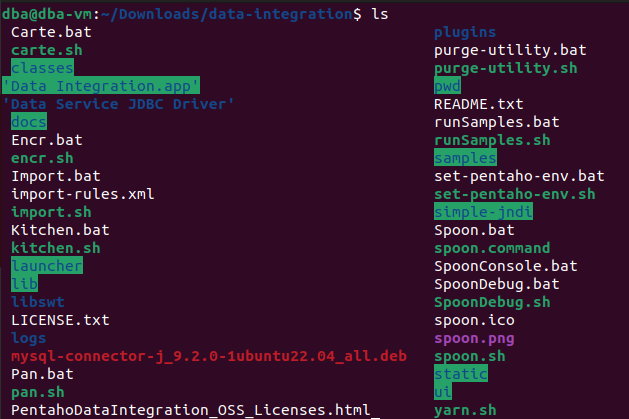


Рисунок 2. Проверка установки коннектора mysql

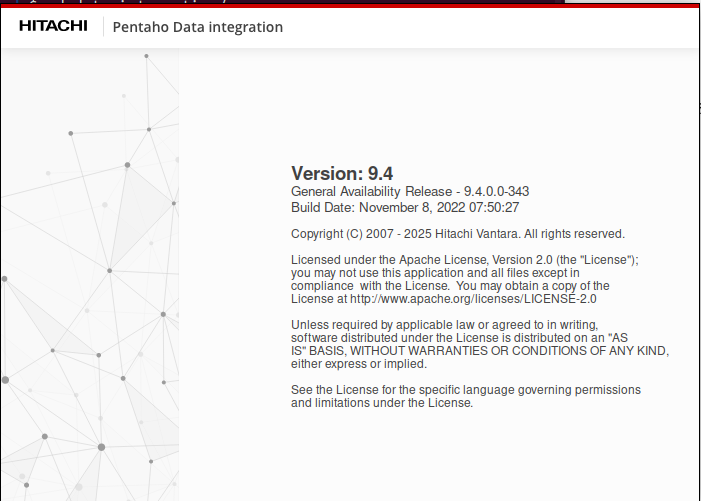


Рисунок 3. Pentaho успешно запущен

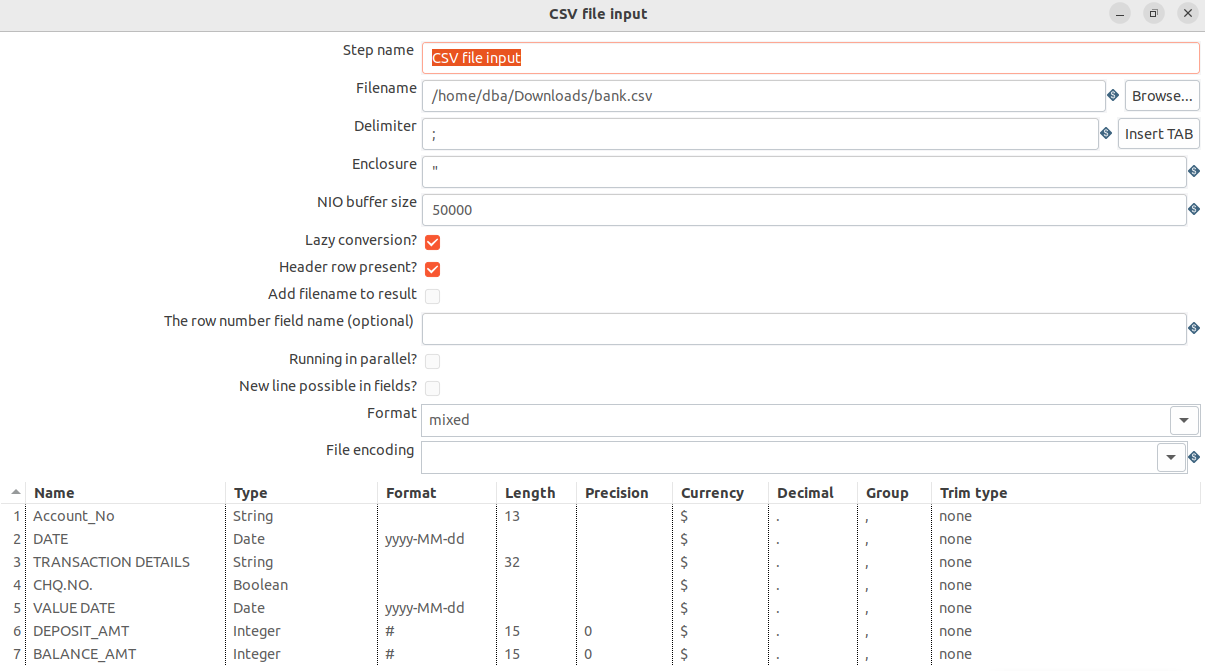


Рисунок 4. Загрузка csv файла

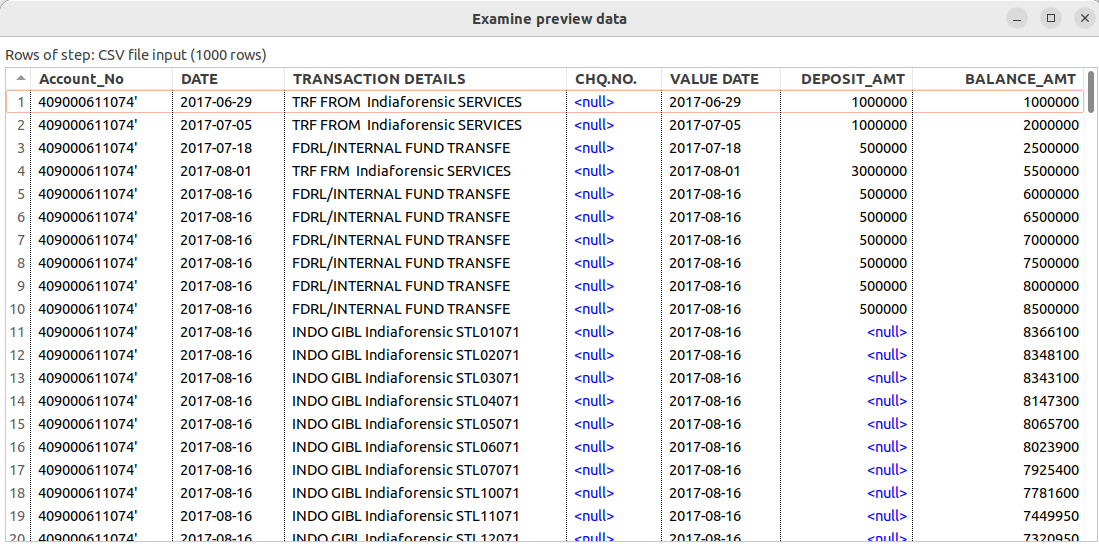


Рисунок 5. Предпросмотр загруженных данных

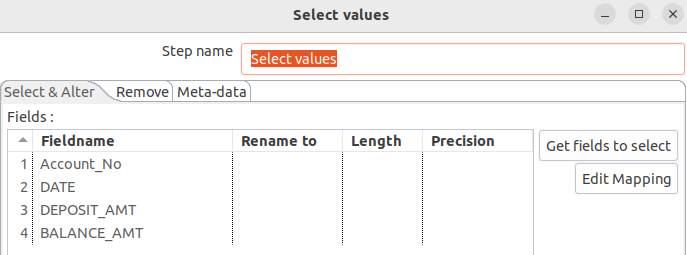


Рисунок 6. Выбор столбцов для дальнейшего анализа

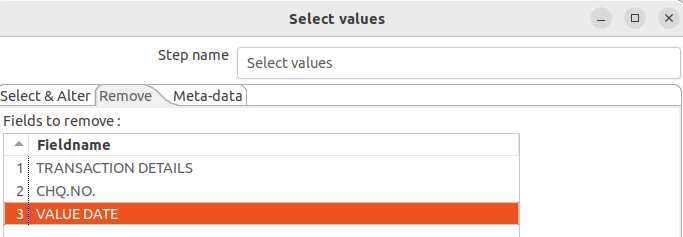


Рисунок 7. Столбцы, выбранные для удаления

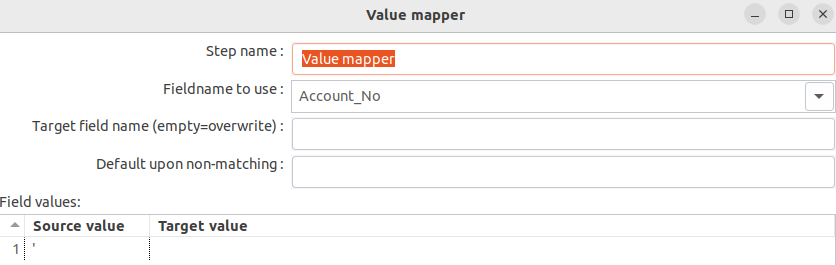


Рисунок 8. Замена лишнего символа в номере аккаунта

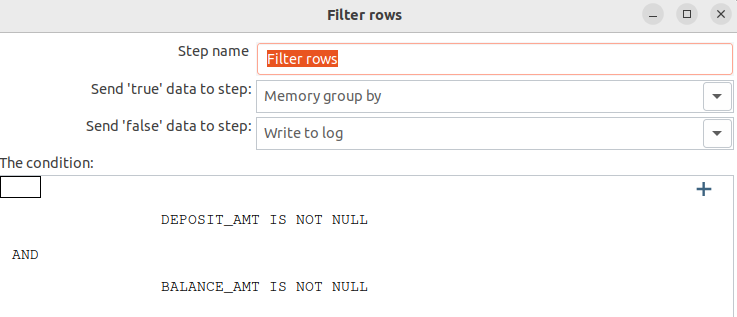


Рисунок 9. Фильтрация нулевых значений в двух столбцах

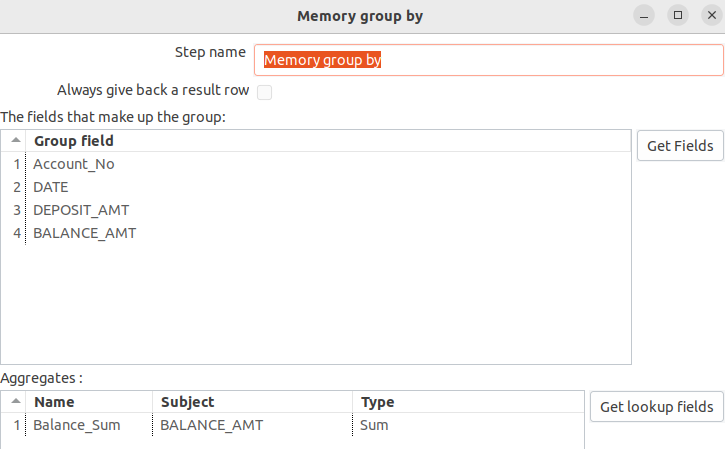


Рисунок 10. Группировка данных по сумме баланса

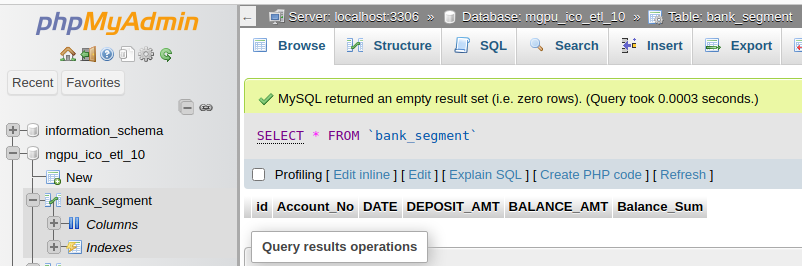


Рисунок 11. Создание таблицы в MySQL

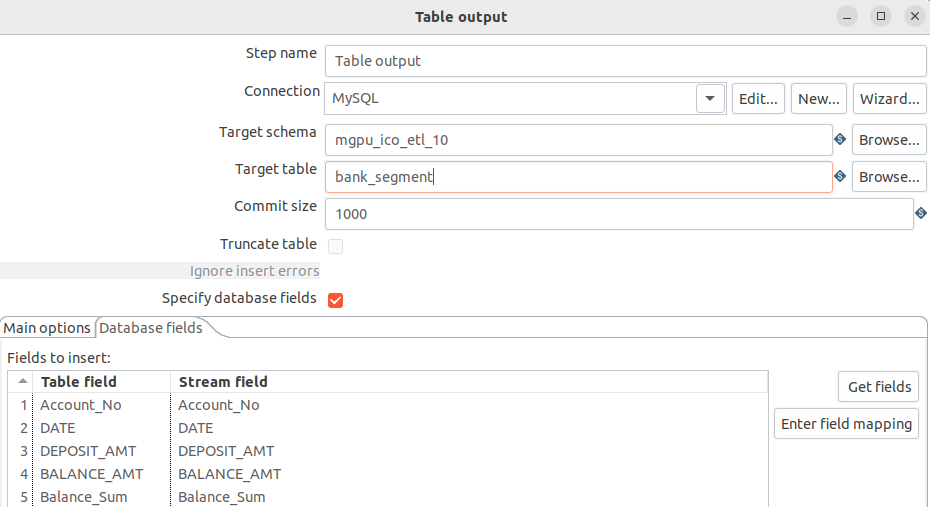


Рисунок 12. Настройка выгрузки данных в бд

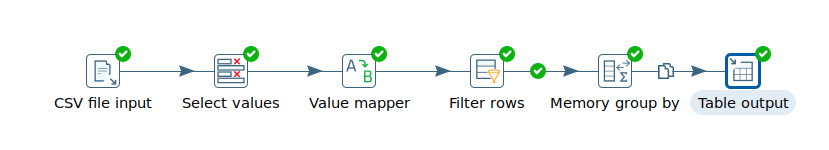


Рисунок 13. Этапы трансформации

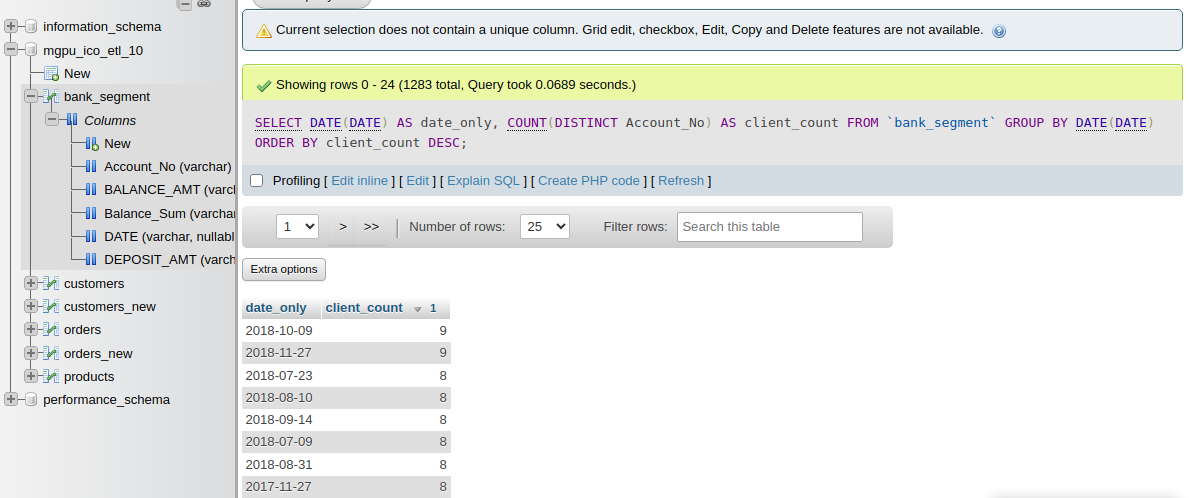


Рисунок 14. Тенденция совершения банковских операций

Как видно, согласно выполненному запросу, большее количество клиентов обращалось в 2018 году в осенний период, приходящийся на октябрь и ноябрь.

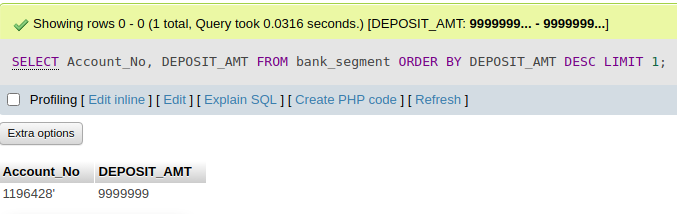


Рисунок 15. Клиент, у которого максимальный депозит

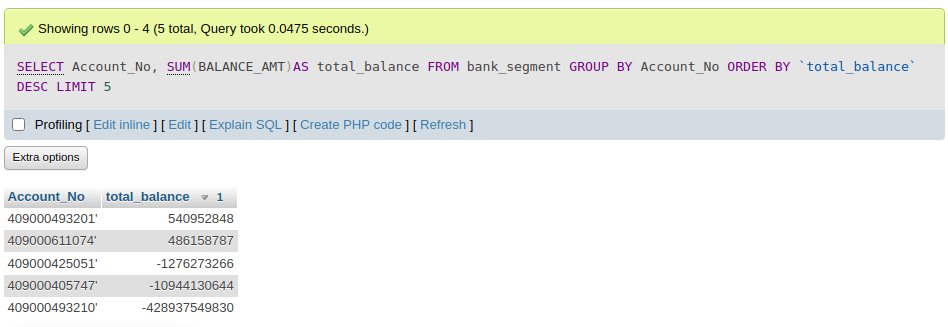


Рисунок 16. Топ 5 клиентов из премиум сегмента, имеющие самый большой баланс

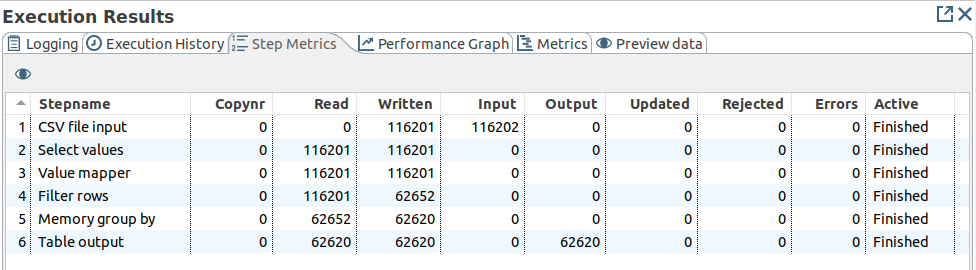


Рисунок 17. Результаты трансформации

Согласно результату, видно, что после группировки количество прочитанных и записанных строк сократилось почти в 2 раза.

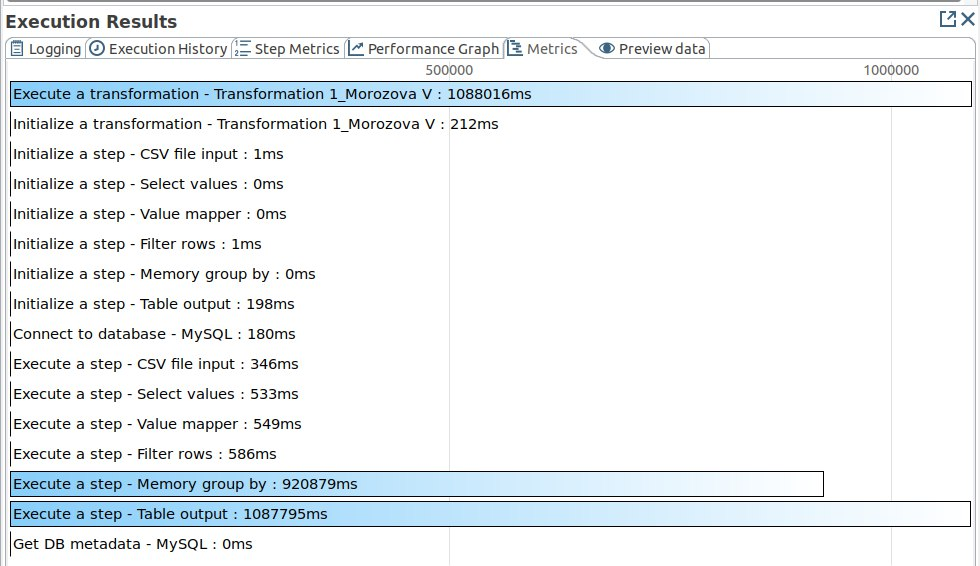


Рисунок 18. Результат выполнения каждого степа по времени в мс

**Выводы:**

1. Было развернуто Pentaho Data Integration.
2. Создан ETL-конвейер:

* загружены данные из CSV-файла
* очищены, преобразованы и отфильтрованы данные
* выполнена замена значений
* выгружены обработанные и сгруппированные данные в MySQL.

1. Проверена корректность обработки.

Выполнены SQL-запросы для проверки результата и сегментации клиентов согласно частоте банковских транзакций.

Согласно результатам, выполнение трансформации заняло 1088016 ms, из которых большую часть заняли процессы группировки и выгрузки в MySQL (920879 и 1087795 ms соответственно).