

Департамент образования города Москвы

Государственное автономное образовательное учреждение  
высшего образования города Москвы  
«Московский городской педагогический университет»

Институт цифрового образования  
Департамент информатики, управления и технологий

**Лабораторная работа 1.1**  
**по дисциплине «Проектный практикум по разработке ETL-решений»**

**Тема: «Установка и настройка ETL-инструмента. Создание конвейеров данных»**

Направление подготовки 38.03.05 – бизнес-информатика  
Профиль подготовки «Аналитика данных и эффективное управление»  
(очная форма обучения)

Выполнила:  
Студентка группы АДЭУ-211  
St\_88

Москва  
2025

## ВВЕДЕНИЕ

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:

1. Настроить среду для работы с Pentaho Data Integration (PDI):
  - Запуск виртуальной машины с Ubuntu 22.04 в VirtualBox.
  - Проверка установки Java и WebKitGTK.
  - Развертывание Pentaho Data Integration.
2. Создать ETL-конвейер:
  - Загрузить данные из CSV-файла.
  - Очистить, преобразовать и отфильтровать данные.
  - Выполнить замену значений.
  - Выгрузить обработанные данные в MySQL или PostgreSQL.
3. Проверить корректность обработки:
  - Выполнить SQL-запросы для проверки результата.
  - Подготовить отчет с описанием проделанных шагов.

Инструменты и технологии:

Компонент	Описание
Ubuntu 22.04 (.ova)	Образ операционной системы для VirtualBox 7.0
VirtualBox 7.0	Виртуализация среды
Pentaho Data Integration 9.4	ETL-инструмент для работы с данными
MySQL/PostgreSQL	База данных для хранения обработанных данных
CSV-файлы	Исходные данные для обработки
Java 11	Необходима для работы Pentaho
libwebkitgtk-1.0-0	Библиотека для корректного запуска Spoon
SQL	Язык запросов для работы с базами данных

**Вариант 1.** Анализ розничных продаж: фильтрация транзакций, выявление аномалий, расчет метрик эффективности

Датасет: Retail Sales Dataset

(<https://www.kaggle.com/datasets/mohammadtali b786/retail-sales-dataset>)

## ХОД РАБОТЫ

1. Pentaho и все необходимые дополнения для его работы были уже установлены на практическом занятии. Поэтому первым шагом стал просто запуск Pentaho:

```
dba@dba-vm: ~/Downloads/data-integration
dba@dba-vm:~$ cd data-integration
bash: cd: data-integration: No such file or directory
dba@dba-vm:~$ cd Downloads
dba@dba-vm:~/Downloads$ cd data-integration
dba@dba-vm:~/Downloads/data-integration$ ./spoon.sh
```

2. Поиск и сохранение датасета Kaggle:

**Retail Sales Dataset** 245 <> Code Download

[Data Card](#) [Code \(52\)](#) [Discussion \(0\)](#) [Suggestions \(0\)](#)

unearth trends, patterns, and correlations that shed light on the complex interplay between customers and products in a retail setting. Happy analyzing!

Transaction ID  
int64

1 1000

Date  
datetime64

2023-01-01 2024-01-01  
11/25/2023 - 01/01/2024  
Count: 106

Customer ID  
object

1000  
unique values

Gender  
object

Female 5  
Male 4

Transaction ID	Date	Customer ID	Gender
1	2023-11-24	CUST001	Male
2	2023-02-27	CUST002	Female

3. Приступаем к созданию трансформации непосредственно в Pentaho.

Для начала, добавляем и настраиваем компонент для импорта данных из CSV-файла:

CSV file input

Step name

CSV file input

Filename

/home/dba/Downloads/retail\_sales\_dataset.csv

Browse...

Delimiter

,

Insert TAB

Enclosure

"

NIO buffer size

50000

Lazy conversion?

☒

Header row present?

☒

Add filename to result

☐

The row number field name (optional)

Running in parallel?

☐

New line possible in fields?

☐

Format

mixed

File encoding

	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Transaction ID	Integer	#	15	0	\$	.	,	none
2	Date	Date	yyyy-MM-dd			\$	.	,	none
3	Customer ID	String		7		\$	.	,	none
4	Gender	String		6		\$	.	,	none
5	Age	Integer	#	15	0	\$	.	,	none
6	Product Category	String		11		\$	.	,	none
7	Quantity	Integer	#	15	0	\$	.	,	none
8	Price per Unit	Integer	#	15	0	\$	.	,	none
9	Total Amount	Integer	#	15	0	\$	.	,	none

Help

OK

Get Fields

Preview

Cancel

В этом же компоненте просматриваем имеющиеся данные:

- **Transaction ID:** Идентификатор транзакции
- **Date:** Дата транзакции
- **Customer ID:** Идентификатор клиента
- **Gender:** Пол клиента
- **Age:** Возраст клиента
- **Product Category:** Категория товара
- **Quantity:** Количество товаров в транзакции
- **Price per Unit:** Цена за единицу товара
- **Total Amount:** Суммарная стоимость транзакции

Examine preview data								
Rows of step: CSV file input (100 rows)								
	Transaction ID	Date	Customer ID	Gender	Age	Product Category	Quantity	Total Amount
1	1	2023-11-24	CUST001	Male	34	Beauty	3	150
2	2	2023-02-27	CUST002	Female	26	Clothing	2	1000
3	3	2023-01-13	CUST003	Male	50	Electronics	1	30
4	4	2023-05-21	CUST004	Male	37	Clothing	1	500
5	5	2023-05-06	CUST005	Male	30	Beauty	2	100
6	6	2023-04-25	CUST006	Female	45	Beauty	1	30
7	7	2023-03-13	CUST007	Male	46	Clothing	2	50
8	8	2023-02-22	CUST008	Male	30	Electronics	4	100
9	9	2023-12-13	CUST009	Male	63	Electronics	2	600
10	10	2023-10-07	CUST010	Female	52	Clothing	4	200
11	11	2023-02-14	CUST011	Male	23	Clothing	2	100
12	12	2023-10-30	CUST012	Male	35	Beauty	3	75
13	13	2023-08-05	CUST013	Male	22	Electronics	3	1500
14	14	2023-01-17	CUST014	Male	64	Clothing	4	120
15	15	2023-01-16	CUST015	Female	42	Electronics	4	2000
16	16	2023-02-17	CUST016	Male	19	Clothing	3	1500
17	17	2023-04-22	CUST017	Female	27	Clothing	4	100
18	18	2023-04-30	CUST018	Female	47	Electronics	2	50
19	19	2023-09-16	CUST019	Female	62	Clothing	2	50
20	20	2023-11-05	CUST020	Male	22	Clothing	3	900
21	21	2023-01-14	CUST021	Female	50	Beauty	1	500
22	22	2023-10-15	CUST022	Male	18	Clothing	2	100
23	23	2023-04-12	CUST023	Female	35	Clothing	4	120
24	24	2023-11-29	CUST024	Female	49	Clothing	1	300
25	25	2023-12-26	CUST025	Female	64	Beauty	1	50
26	26	2023-10-07	CUST026	Female	28	Electronics	2	1000
27	27	2023-08-03	CUST027	Female	38	Beauty	2	50

#### 4. Вывод и переименование полей:

Select values

Step name

Select & Alter

Remove

Meta-data

Fields to alter the meta-data for :

	Fieldname	Rename to	Type	Length	Precision
1	Transaction ID	transaction_id	Integer	15	0
2	Date	date	Date		
3	Customer ID	customer_id	Integer	7	
4	Gender	gender	String	6	
5	Age	age	Integer	15	0
6	Product Category	product_category	String	11	
7	Quantity	quantity	Integer	15	0
8	Price per Unit	price_per_unit	Integer	15	0
9	Total Amount	total_amount	Integer	15	0

Get fields to change

5. Фильтрация данных от нулевых значений идентификатора транзакции и даты транзакции:

Filter rows

Step name:

Send 'true' data to step:

Send 'false' data to step:

The condition:

☐ `transaction_id IS NOT NULL`

AND

`date IS NOT NULL`

6. Приведения данных о поле клиента к булевым значениям 1 – женщины, 0 – мужчины:

Value mapper

Step name:

Fieldname to use:

Target field name (empty=overwrite):

Default upon non-matching:

Field values:

	Source value	Target value
1	Female	1
2	Male	0

7. Сортировка данных по категориям товаров и полу клиентов (подготовка к дальнейшей группировке):

Sort rows

Step name

Sort rows 2

Sort directory

%%java.io.tmpdir%%

Browse...

TMP-file prefix

out

Sort size (rows in memory)

1000000

Free memory threshold (in %)

Compress TMP Files?

☐

Only pass unique rows? (verifies keys only)

☐

Fields :

	Fieldname	Ascending	Case sensitive compare?	Sort based on current locale?	Collator Strength	Presorted
1	product_category	Y	N	N	0	N
2	gender	Y	N	N	0	N

8. Группировка данных по категориям товаров и полу клиентов. Создание новых агрегированных полей:

- Общее количество транзакций;
- Общее количество купленных товаров;
- Итоговая сумма транзакций (д.е.);
- Минимальная сумма транзакции (д.е.);
- Максимальная сумма транзакции (д.е.);
- Средняя сумма транзакции (д.е.);
- Медианная сумма транзакции (д.е.)

**Group by**

Step name:

Include all rows? ☐

Temporary files directory:

TMP-file prefix:

Add line number, restart in each group ☐

Line number field name:

Always give back a result row ☐

The fields that make up the group:

▲ **Group field**

1	product_category
2	gender

Aggregates :

	Name	Subject	Type	Value
1	total_amount_transactions	transaction_id	Number of Values (N)	
2	total_quantity	quantity	Sum	
3	total_sales	total_amount	Sum	
4	max_sales	total_amount	Maximum	
5	min_sales	total_amount	Minimum	
6	avarege_sales	total_amount	Average (Mean)	
7	median_sales	total_amount	Median	

9. Подготовка к экспорту данных: на сервере MySQL создаем таблицу transactions с помощью следующего скрипта:

```

DROP TABLE IF EXISTS transactions;

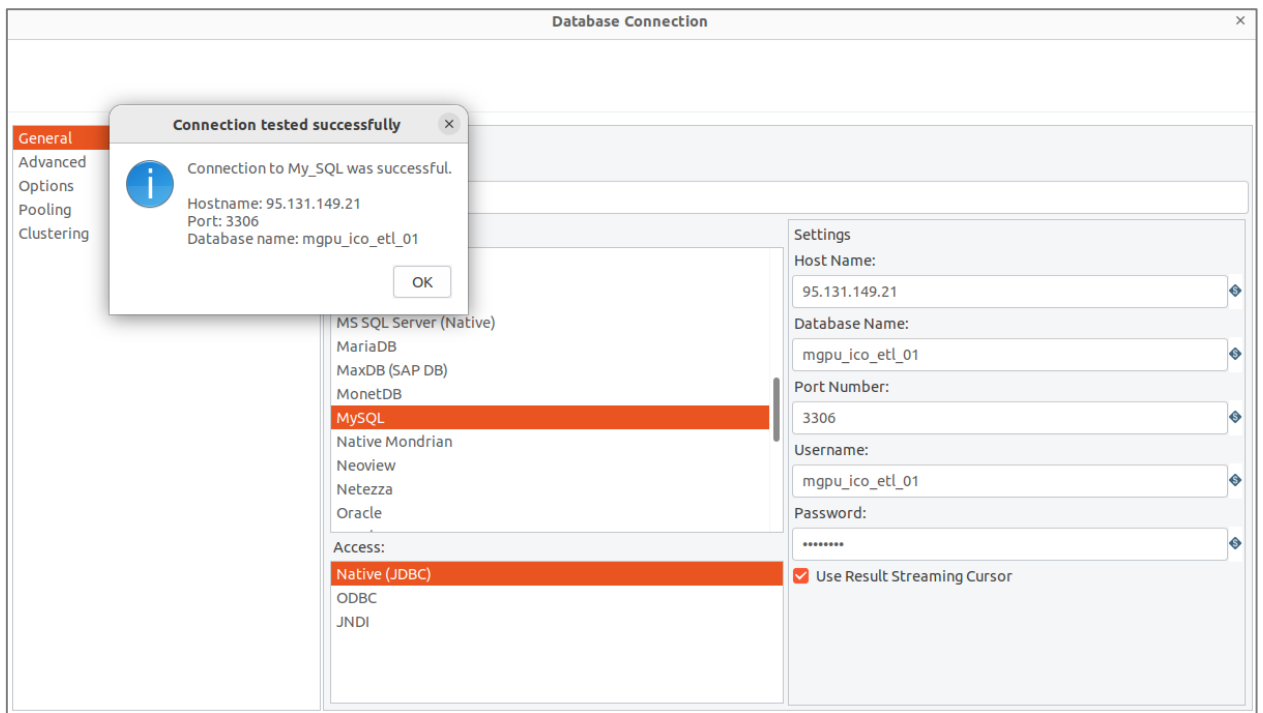
CREATE TABLE transactions (
  id INT AUTO_INCREMENT PRIMARY KEY,
  gender TINYINT(1) DEFAULT 0,
  product_category VARCHAR(100),
  total_amount_transactions INT,
  total_quantity INT,
  total_sales INT,
  max_sales INT,
  min_sales INT,
  avarege_sales INT,
  median_sales INT
);

```

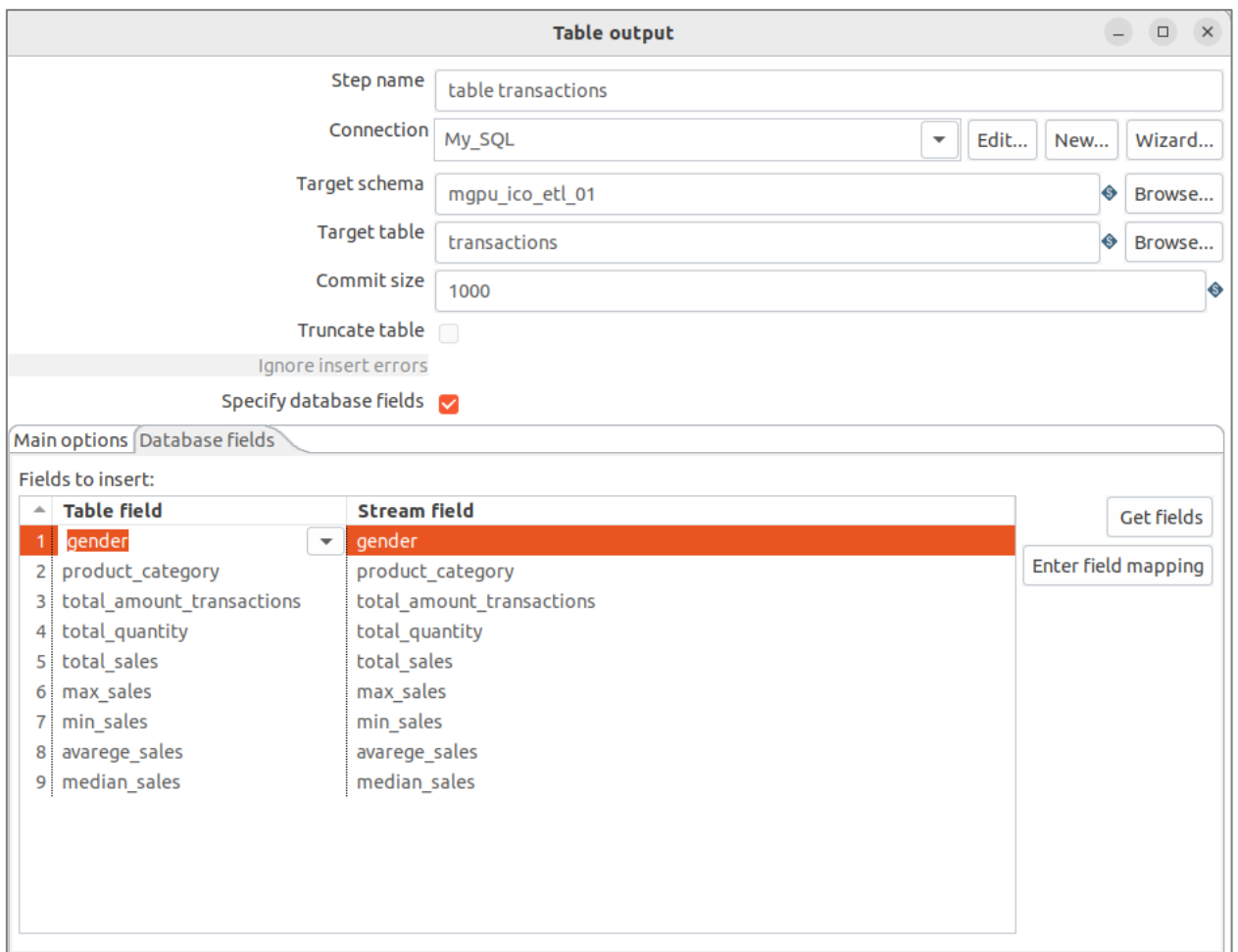
10. Добавление компонента экспорта результата в базу данных.

Настраиваем подключение к базе данных MySQL:

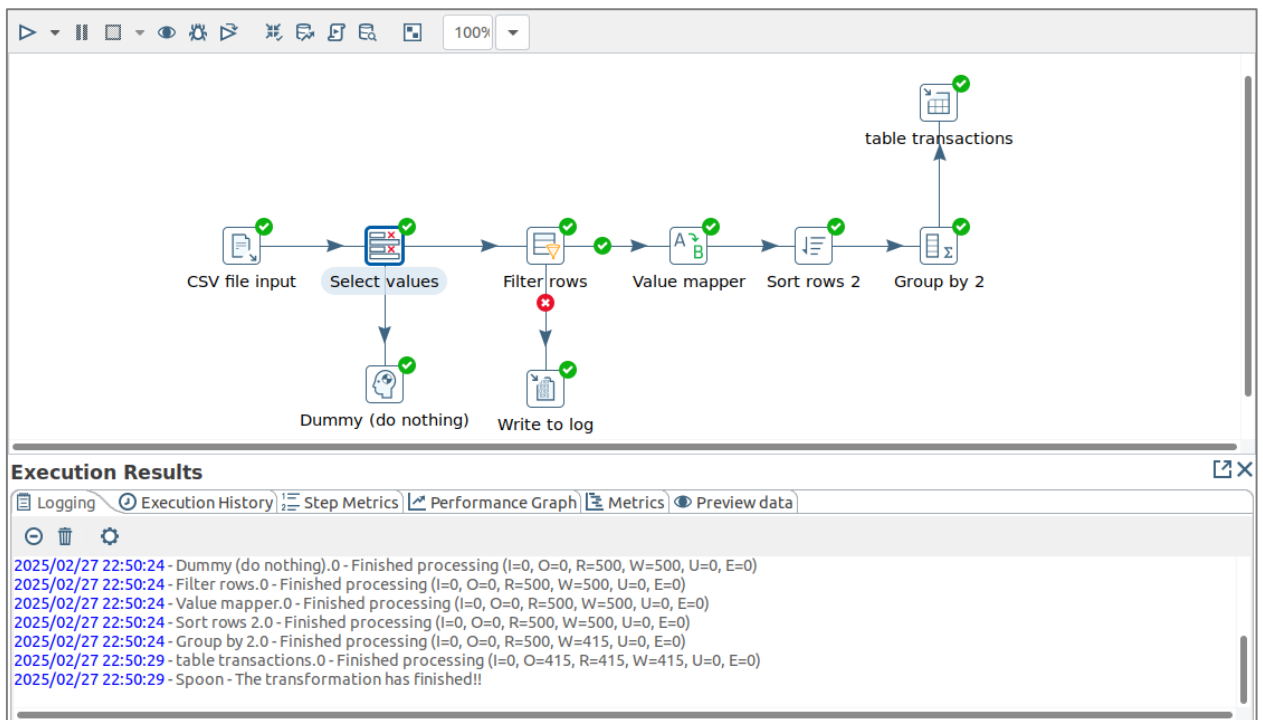




Выбор необходимой таблицы из базы данных и маппинг полей:



11. В итоге получаем следующую трансформацию:



12. Запускаем трансформацию и имеем следующий результат в базе данных:

Отображение строк 0 - 5 (6 всего, Запрос занял 0.0002 сек.) [gender: 0... - 1...]

`SELECT * FROM `transactions` ORDER BY `transactions`.`gender` ASC`

☐ Профилирование ☐ Построчное редактирование ☐ Изменить ☐ Анализ SQL запроса ☐ Создать PHP-код ☐ Обновить

☐ Показать все | Количество строк: 25 | Фильтровать строки: Поиск в таблице | Сортировать по ключу: Ни одного

Extra options

	id	gender	product_category	total_amount_transactions	total_quantity	total_sales	max_sales	min_sales	avarege_sales	median_sales
<input type="checkbox"/> Изменить <input type="checkbox"/> Копировать <input type="checkbox"/> Удалить	1	0	Beauty	60	156	27005	2000	25	450	120
<input type="checkbox"/> Изменить <input type="checkbox"/> Копировать <input type="checkbox"/> Удалить	3	0	Clothing	92	249	41830	2000	25	454	150
<input type="checkbox"/> Изменить <input type="checkbox"/> Копировать <input type="checkbox"/> Удалить	5	0	Electronics	80	189	37475	2000	25	468	135
<input type="checkbox"/> Изменить <input type="checkbox"/> Копировать <input type="checkbox"/> Удалить	2	1	Beauty	91	232	39840	2000	25	437	150
<input type="checkbox"/> Изменить <input type="checkbox"/> Копировать <input type="checkbox"/> Удалить	4	1	Clothing	96	236	48145	2000	25	501	135
<input type="checkbox"/> Изменить <input type="checkbox"/> Копировать <input type="checkbox"/> Удалить	6	1	Electronics	81	208	35415	2000	25	437	150

☐ Отметить все | С отмеченными: ☐ Изменить ☐ Копировать ☐ Удалить ☐ Экспорт

☐ Показать все | Количество строк: 25 | Фильтровать строки: Поиск в таблице | Сортировать по ключу: Ни одного

Выводы, которые можно сделать из данного результата:

Женщины произвели транзакции на большую сумму и большее количество, чем мужчины.

Наиболее популярная категория как среди мужчин, так и среди женщин – одежда.

Мужчины произвели больше всего покупок в категории одежды, общая стоимость транзакций по данной категории среди мужчин составила 41 830

д.е., при том что средний чек составил 454 д.е. (наиболее частый чек – 150 д.е.). На втором месте категория электроники, на третьем – товары для красоты.

Женщины также произвели больше всего транзакций по категории «Одежда», общая стоимость которых составила 48 145 д.е., при том что средний чек составил 501 д.е. (наиболее частый чек – 135 д.е.). На втором месте категория товаров для красоты, на третьем – электроника.

## ИТОГИ

В ходе выполнения данной лабораторной работы были реализованы все поставленные задачи:

1. Была настроена среда для работы и успешно запущено ПО Pentaho Data Integration;
2. Был создан ETL-конвейер с загрузкой данных из CSV-файла, фильтрацией, преобразованием и группировкой данных, а также с экспортом данных в MySQL;
3. Были проанализированные полученные результаты рассмотрения данных о транзакциях и сформированы выводы.

Таким образом, была достигнута цель работы: изучены основные принципы работы с ETL-инструментами на примере Pentaho Data Integration (PDI).