

Департамент образования 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»

Выполнила:

Студентка группы АДЭУ-211

Кравцова Алёна Евгеньевна

Руководитель:

Босенко Т.М

Москва

2024

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.

Индивидуальное задание: Вариант 6. Маркетинговая аналитика: анализ эффективности рекламных кампаний

Шаги выполнения:

1. Запуск pentaho

Ввиду того, что работа осуществляется в подготовленном образе шаги по установке pentaho отсутствуют. Запустим pentaho (Рис. 1).

```
dba@dba-vm:~/Downloads/data-integration$ chmod +x spoon.sh
dba@dba-vm:~/Downloads/data-integration$ ./spoon.sh
```

Рис. 1 – Запуск pentaho

В итоге открылось pentaho, запуск произошел успешно (Рис. 2).

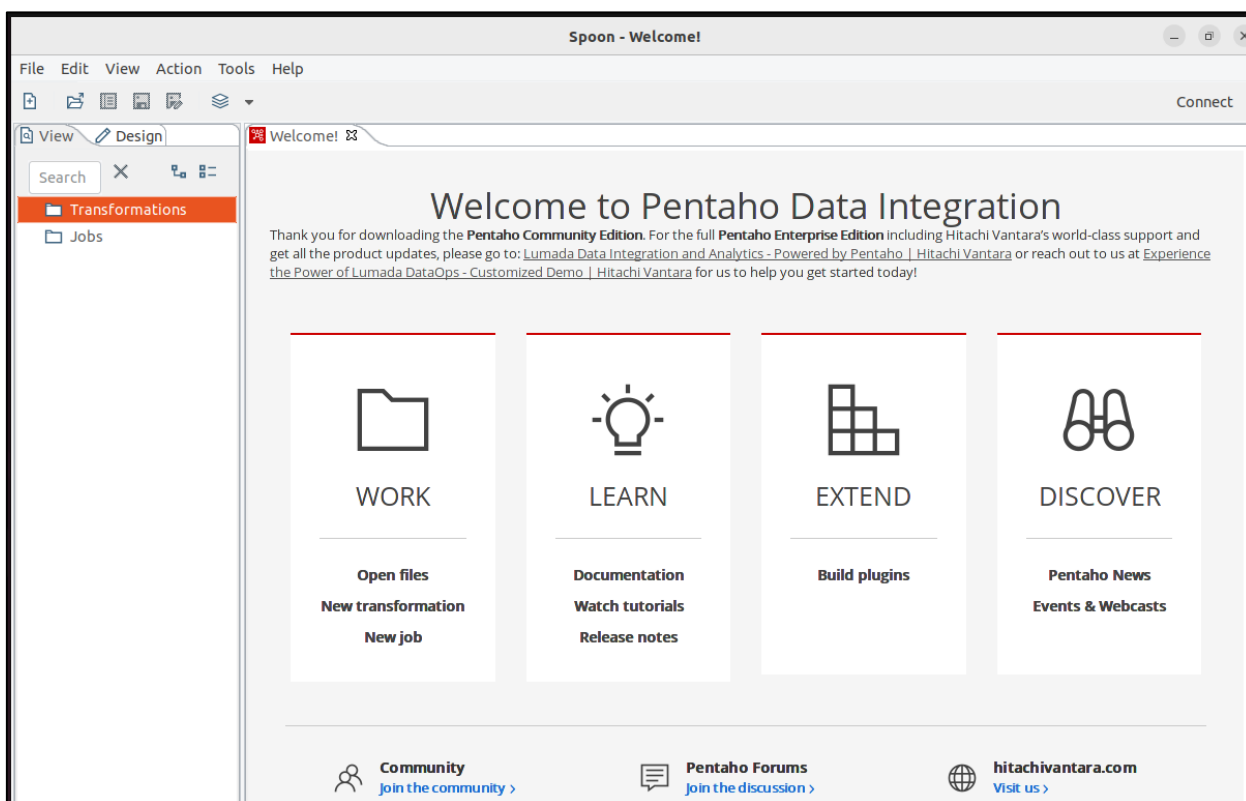


Рис. 2 – Стартовая страница pentaho

2. Установка необходимых разрешений, для подключения к БД

Для подключения к БД необходимо загрузить my sql connector (Рис. 3).

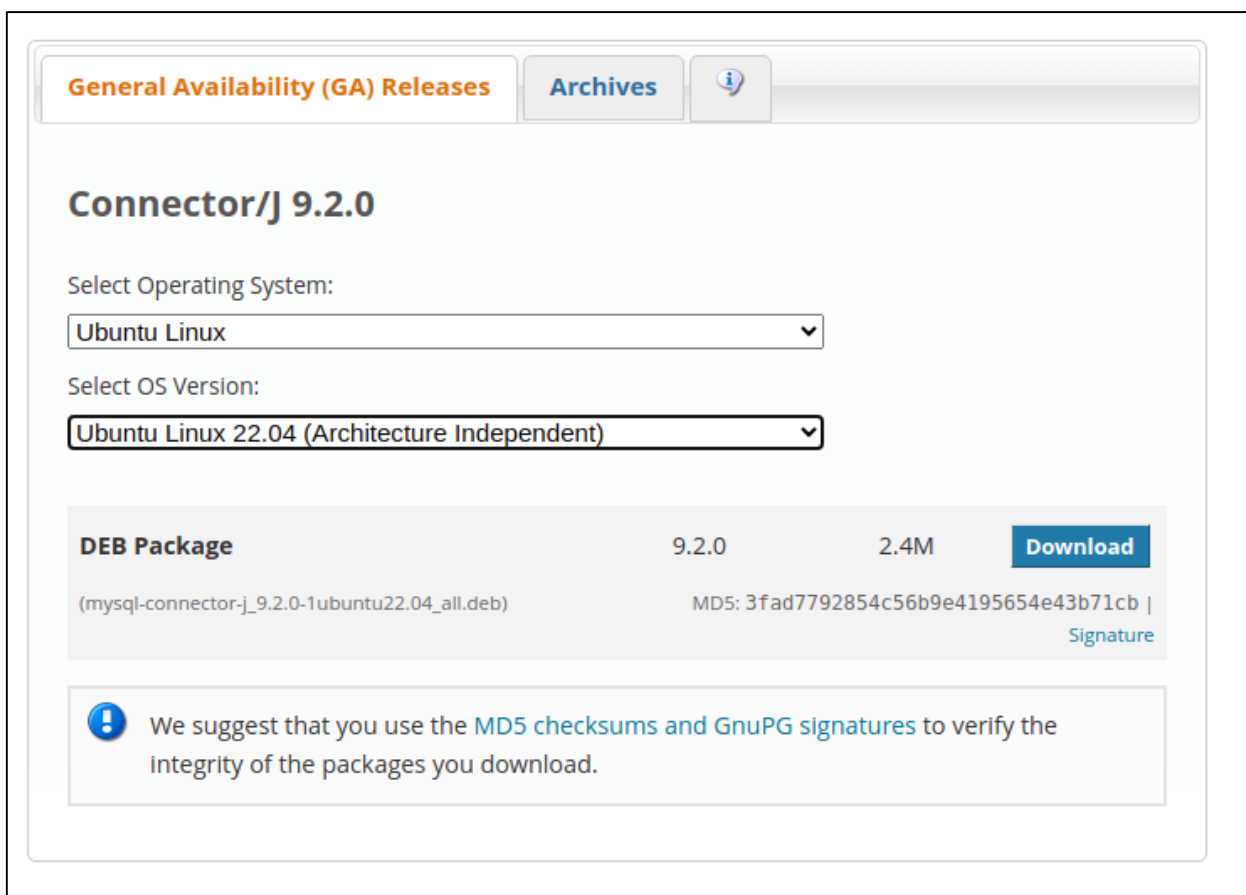


Рис. 3 – Загрузка my sql connector

После загрузки разархивируем пакет и перенесем необходимый файл в папку data integration (Рис. 4, Рис. 5).

```
dba@dba-vm:~/Downloads$ sudo dpkg -i mysql-connector-j_9.2.0-1ubuntu22.04_all.deb
Selecting previously unselected package mysql-connector-j.
(Reading database ... 236562 files and directories currently installed.)
Preparing to unpack mysql-connector-j_9.2.0-1ubuntu22.04_all.deb
...
Unpacking mysql-connector-j (9.2.0-1ubuntu22.04) ...
Setting up mysql-connector-j (9.2.0-1ubuntu22.04) ...
```

Рис. 4 – Разархивация загруженной папки

```
dba@dba-vm:~/Downloads$ ls /usr/share/java/mysql-connector-j-9.2.0.jar
/usr/share/java/mysql-connector-j-9.2.0.jar
dba@dba-vm:~/Downloads$ sudo cp /usr/share/java/mysql-connector-j-9.2.0.jar ~/Downloads/data-integration/lib
```

Рис. 5 – Перенос файла в папку с pentaho

3. Индивидуальное задание

Для анализа данных маркетинговых компаний откроем соответствующий датасет Marketing Campaign на kaggle.

Создадим новую трансформацию, в которую необходимо добавить узел с загрузкой csv файла. Далее в настройках компонента выбрать путь к файлу и настроить разделитель (Рис. 6).

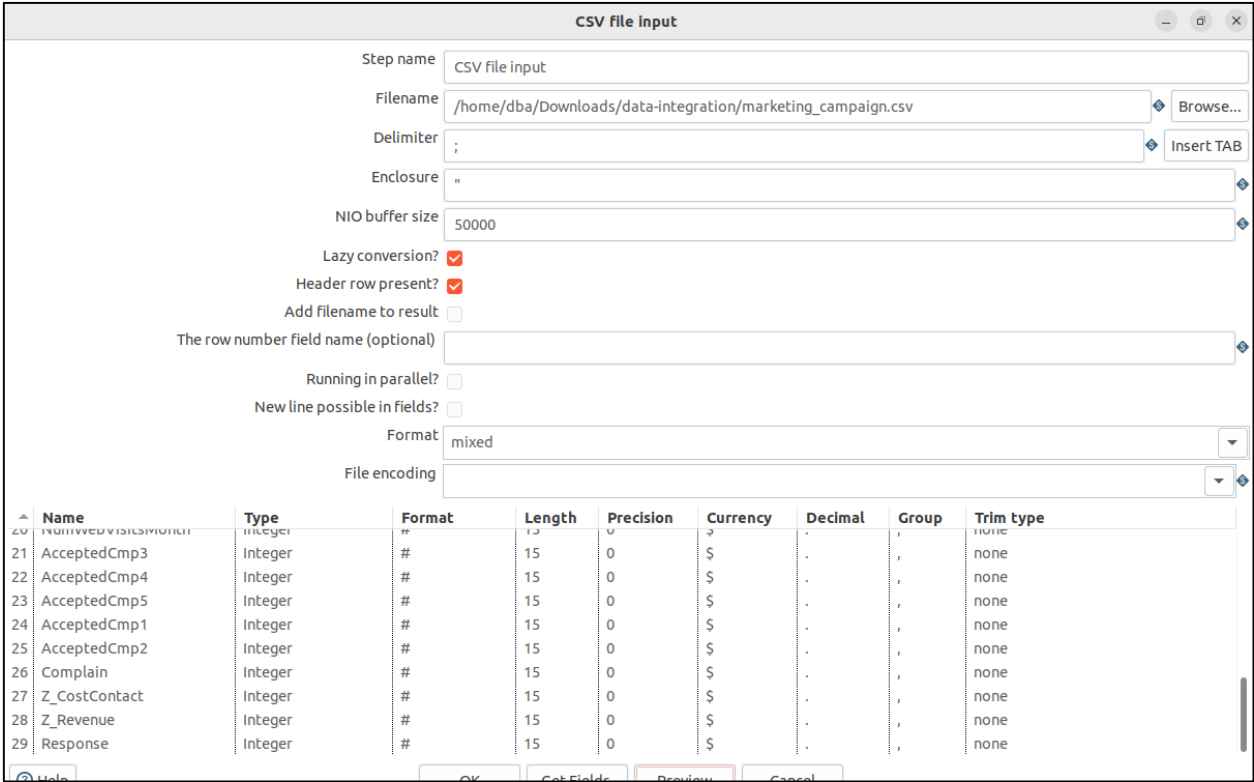


Рис. 6 – Csv file input

Рассмотрим загруженный датасет внимательнее. Описание полей представлено в таблице 1.

Таблица 1

Название поля	Описание	Тип данных
ID	Идентификатор клиента	integer
Year_birth	Год рождения клиента	integer
Education	Уровень образования клиента	string
Marital_Status	Семейное положение клиента	string
Income	Доход	integer
Kidhome	Количество маленьких детей в семье клиента	integer

Teenhome	Количество подростков в семье клиента	integer
Dt_customer	Дата регистрации клиента в компании	date
Recency	Количество дней с момента последней покупки	integer
MntFishProducts	Сумма, потраченная на рыбные продукты за последние 2 года	integer
MntMeatProducts	Сумма, потраченная на мясные продукты за последние 2 года	integer
MntFruits	Сумма, потраченная на фрукты за последние 2 года	integer
MntSweetProducts	Сумма, потраченная на сладкое за последние 2 года	integer
MntWines	Сумма, потраченная на вино за последние 2 года	integer
MntGoldProds	Сумма, потраченная на золотые изделия за последние 2 года	integer
NumDealsPurchases	Количество покупок, совершенных со скидкой	integer
NumCatalogPurchases	Количество покупок, совершенных с использованием каталога	integer
NumStorePurchases	Количество покупок, совершенных непосредственно в магазинах	integer
NumWebPurchases	Количество покупок, совершенных онлайн	integer
NumWebVisitsMonth	Количество посещений веб-сайта компании за последний месяц	integer
AcceptedCmp1	1, если клиент принял предложение в рамках 1-й кампании, в противном случае 0	integer
AcceptedCmp2	1, если клиент принял предложение в рамках 2-й кампании, в противном случае 0	integer
AcceptedCmp3	1, если клиент принял предложение в рамках 3-й кампании, в противном случае 0	integer

AcceptedCmp4	1, если клиент принял предложение в рамках 4-й кампании, в противном случае 0	integer
AcceptedCmp5	1, если клиент принял предложение в рамках 5-й кампании, в противном случае 0	integer
Response	1, если клиент принял предложение в последней кампании, в противном случае 0	integer
Complain	1, если клиент жаловался в течение последних 2 лет	integer

Далее добавим компонент Select values для отбора необходимых для анализа данных, во вкладке «remove» необходимо перечислить столбцы, которые не будут использоваться (Рис. 7).

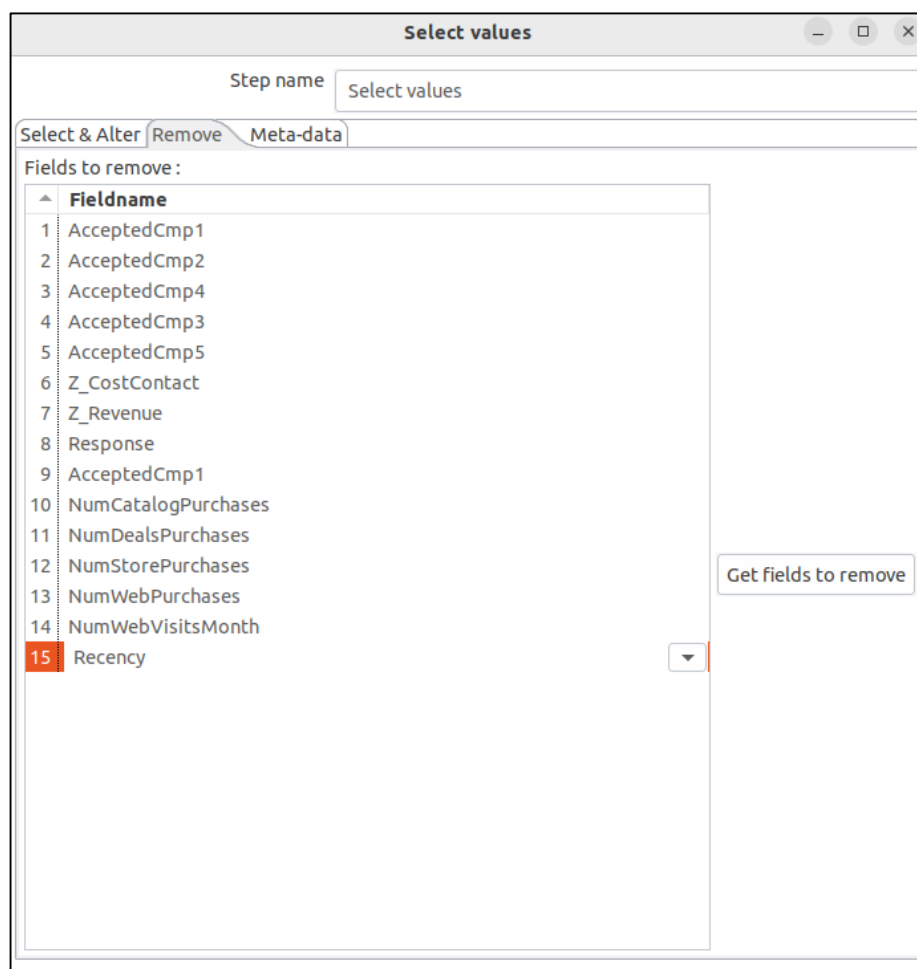


Рис. 7 – Очистка неиспользуемых полей

Также при просмотре данных было обнаружено, что столбец income содержит столбцы без данных (Рис. 8).

Examine preview data												
Rows of step: CSV file input (1000 rows)												
	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	MntFruits	Mi
10	5899	1950	PNL	together	5648	1	1	2014-03-13	68	28	0	
11	1994	1983	Graduation	Married	<null>	1	0	2013-11-15	11	5	5	
12	387	1976	Basic	Married	7500	0	0	2012-11-13	59	6	16	
13	2125	1959	Graduation	Divorced	63033	0	0	2013-11-15	82	194	61	
14	8180	1952	Master	Divorced	59354	1	1	2013-11-15	53	233	2	
15	2569	1987	Graduation	Married	17323	0	0	2012-10-10	38	3	14	
16	2114	1946	PhD	Single	82800	0	0	2012-11-24	23	1006	22	
17	9736	1980	Graduation	Married	41850	1	1	2012-12-24	51	53	5	
18	4939	1946	Graduation	Together	37760	0	0	2012-08-31	20	84	5	
19	6565	1949	Master	Married	76995	0	1	2013-03-28	91	1012	80	
20	2278	1985	2n Cycle	Single	33812	1	0	2012-11-03	86	4	17	
21	9360	1982	Graduation	Married	37040	0	0	2012-08-08	41	86	2	
22	5376	1979	Graduation	Married	2447	1	0	2013-01-06	42	1	1	
23	1993	1949	PhD	Married	58607	0	1	2012-12-23	63	867	0	
24	4047	1954	PhD	Married	65324	0	1	2014-01-11	0	384	0	
25	1409	1951	Graduation	Together	40689	0	1	2013-03-18	69	270	3	
26	7892	1969	Graduation	Single	18589	0	0	2013-01-02	89	6	4	
27	2404	1976	Graduation	Married	53359	1	1	2013-05-27	4	173	4	
28	5255	1986	Graduation	Single	<null>	1	0	2013-02-20	19	5	1	
29	9422	1989	Graduation	Married	38360	1	0	2013-05-31	26	36	2	
30	1966	1965	PhD	Married	84618	0	0	2013-11-22	96	684	100	
31	6864	1989	Master	Divorced	10979	0	0	2014-05-22	34	8	4	
32	3033	1963	Master	Together	38620	0	0	2013-05-11	56	112	17	
33	5710	1970	Graduation	Together	40548	0	1	2012-10-10	31	110	0	
34	7373	1952	PhD	Divorced	46610	0	2	2012-10-29	8	96	12	
35	8755	1946	Master	Married	68657	0	0	2013-02-20	4	482	34	
36	10738	1951	Master	Single	49389	1	1	2013-08-29	55	40	0	

Рис. 8 – Предпросмотр данных

Таким образом, необходимо исключить пустые строки посредством компонента filter rows (Рис. 9).

Filter rows

Step name

Filter rows

Send 'true' data to step:

Send 'false' data to step:

The condition:

Income

IS NOT NULL

Help

OK

Cancel

Рис. 9 – Исключение нулевых значений

Далее необходимо высчитать и записать в новый столбец общие траты клиентов на все категории продукции. В связи с тем, что необходимо сложить 6 столбцов, а pentaho ограничено расчетом с использованием трех столбцов, то в работе будет 2 калькулятора: первый для подсчета промежуточных значений по существующим данным, а второй для создания столбца на основе «новых» столбцов, созданных в калькуляторе 1.

Итак в первом калькуляторе высчитываем промежуточные значения (Рис. 10).

	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove
1	fish_meat_fruits	A + B + C	MntFishProducts	MntMeatProducts	MntFruits	Integer			N
2	sweet_wine_gold	A + B + C	MntSweetProducts	MntWines	MntGoldProds	Integer			N

Рис. 10 – Создание промежуточных полей

Во втором калькуляторе высчитываем итоговую сумму покупок (Рис. 11).

Calculator

Step name

Calculator 2

☒ Throw an error on non existing files

Fields:

	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove	Conversion mask
1	total_spend	A + B	fish_meat_fruits	sweet_wine_gold		Integer			N	

Help

OK

Cancel

Рис. 11 – Расчет итоговой суммы расходов клиента

В связи с тем, что промежуточные вычисления не нужны в БД, то необходимо еще раз отфильтровать данные (Рис. 12).

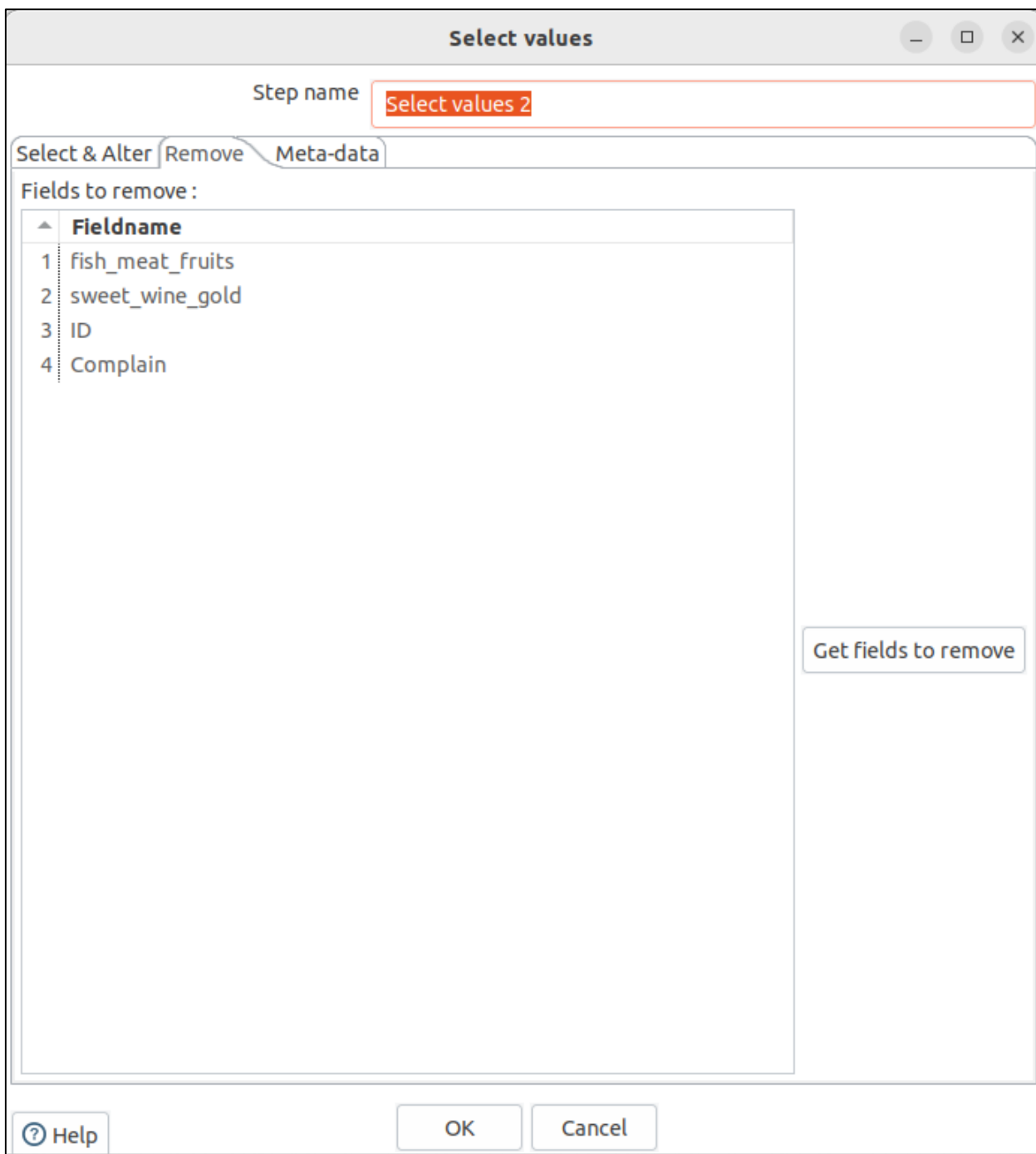


Рис. 12 – Отсеивание неиспользуемых значений

Далее добавим компонент, отвечающий за экспорт данных в БД. Необходимо настроить подключение к базе и проверить его (Рис. 13).

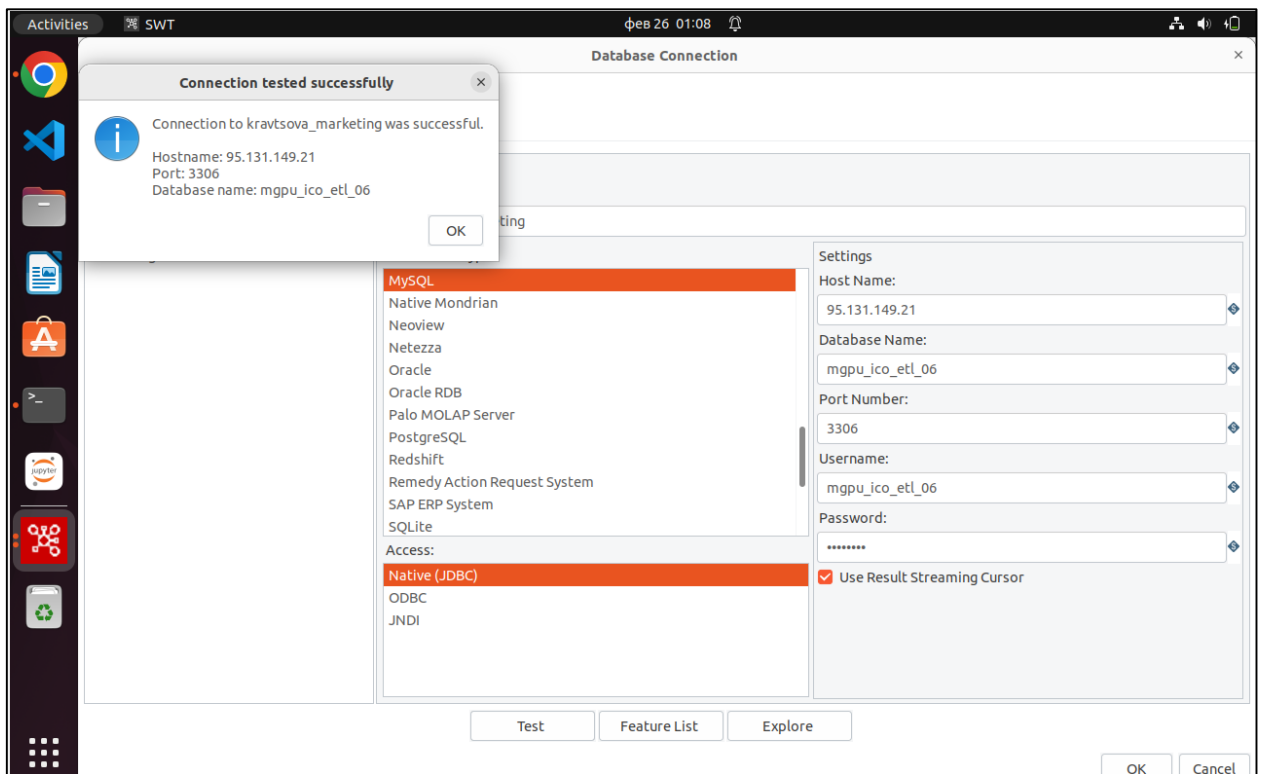


Рис. 13 – Настройка и проверка подключения к БД

Далее необходимо зайти в php admin, в свою учетную запись, и создать таблицу для дальнейшей загрузки данных в нее (Рис. 14).

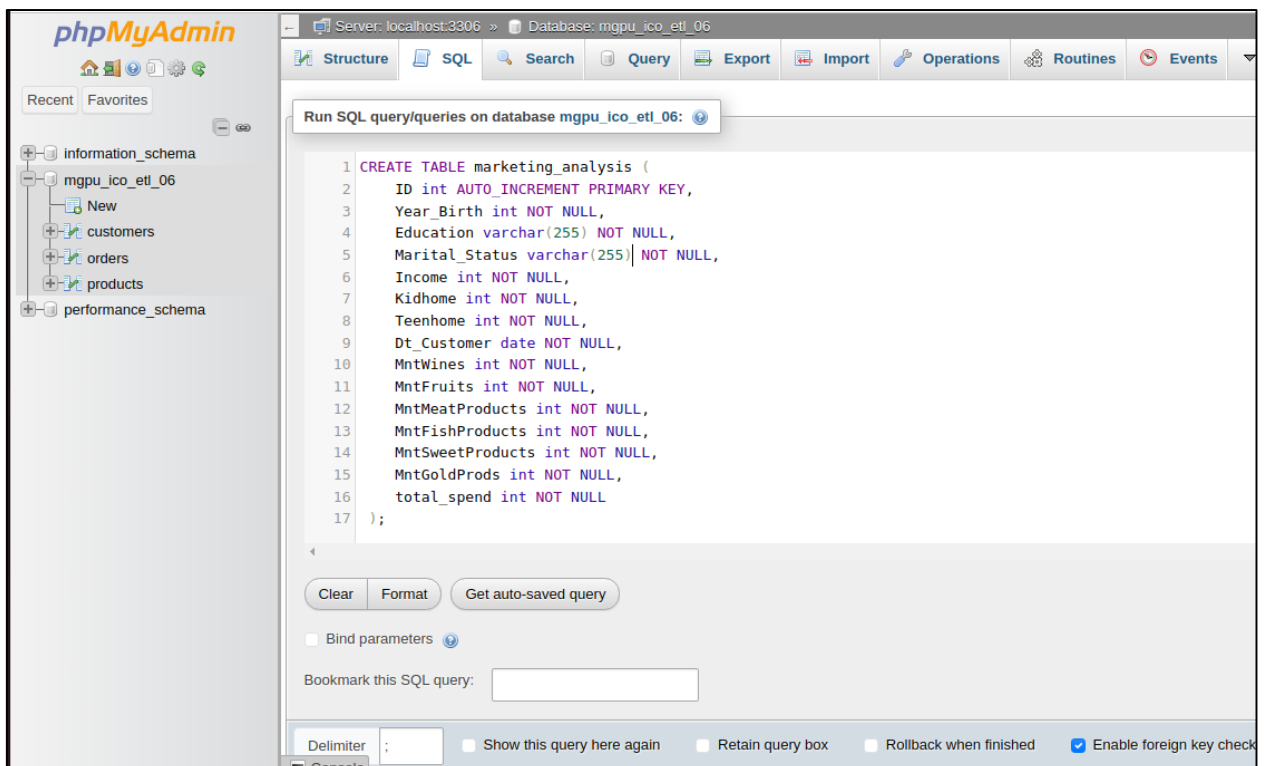


Рис. 14 – Запрос по созданию таблицы

Таблица успешно добавлена (Рис. 15).

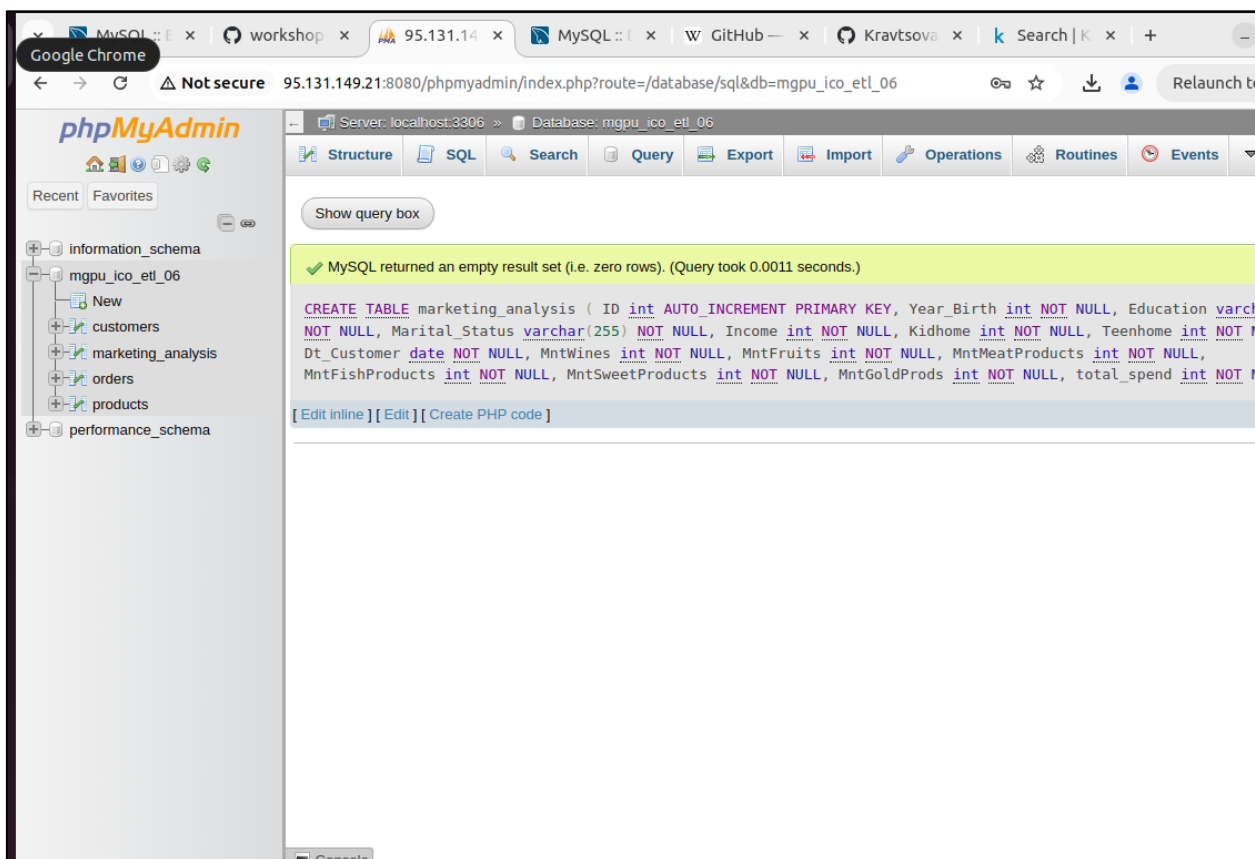


Рис. 15 – Успешное создание таблицы в БД

Укажем таблицу, в которой необходимо записать данные, а также проверим на соответствие перечень передаваемых данных и перечень столбцов из базы (Рис. 16).

Table output

Step name:

Connection:

Target schema:

Target table:

Commit size:

Truncate table: ☐

Ignore insert errors: ☐

Specify database fields: ☒

Main options | **Database fields**

Fields to insert:

	Table field	Stream field
1	Year_Birth	Year_Birth
2	Education	Education
3	Marital_Status	Marital_Status
4	Income	Income
5	Kidhome	Kidhome
6	Teenhome	Teenhome
7	Dt_Customer	Dt_Customer
8	MntWines	MntWines
9	MntFruits	MntFruits
10	MntMeatProd	MntMeatProducts
11	MntFishProdu	MntFishProducts
12	MntSweetPro	MntSweetProducts
13	MntGoldProd	MntGoldProds
14	total_spend	total_spend

Рис. 16 – Проверка соответствия столбцов выбранной таблицы

Далее выполним трансформацию, трансформация выполнена успешно, pentaho не выдал никаких ошибок (Рис. 17).

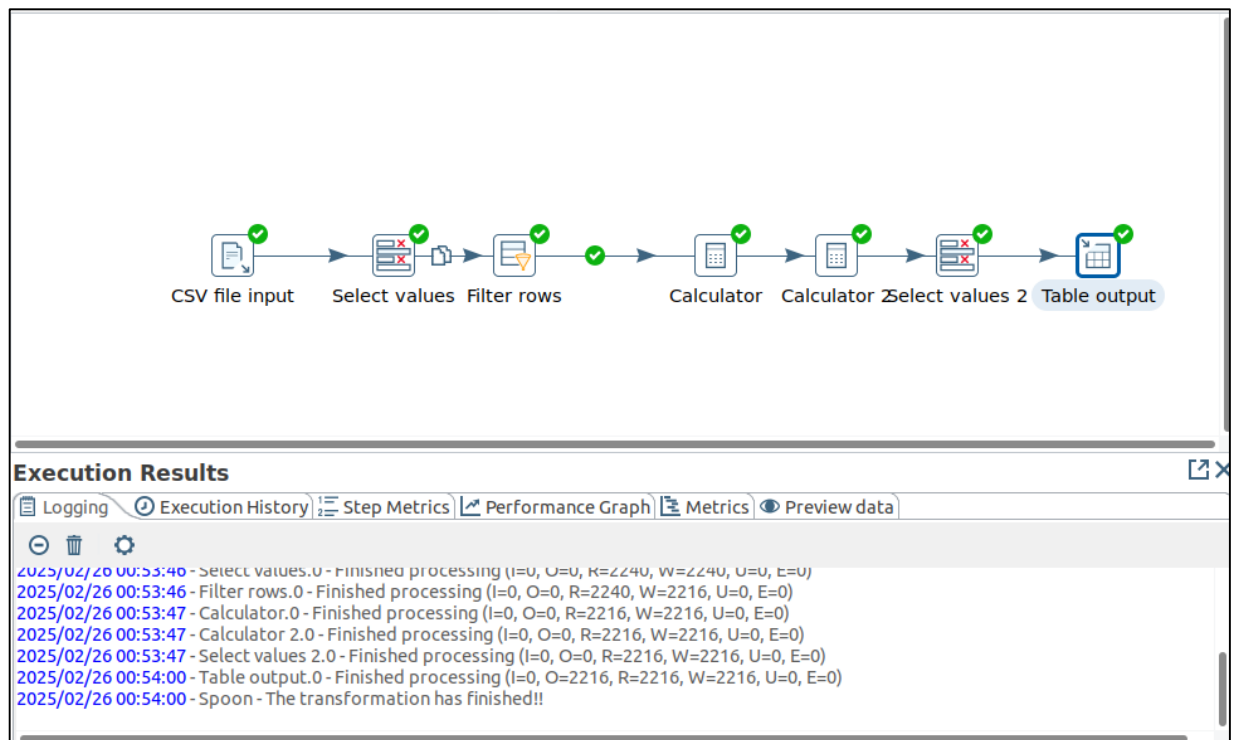


Рис. 17 – Выполнение трансформации в pentaho

Перейдем в php admin, чтобы убедиться, что все данные загрузились корректно (Рис. 18).

Not secure 95.131.149.21:8080/phpmyadmin/index.php?route=/sql&pos=0&db=mgpu_ico_etl_06&table=ma... Relaunch to update

Server: localhost3306 » Database: mgpu_ico_etl_06 » Table: marketing_analysis

Showing rows 0 - 24 (2216 total. Query took 0.0002 seconds.)

SELECT * FROM `marketing_analysis`

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> Number of rows: 25 Filter rows: Search this table Sort by key: None

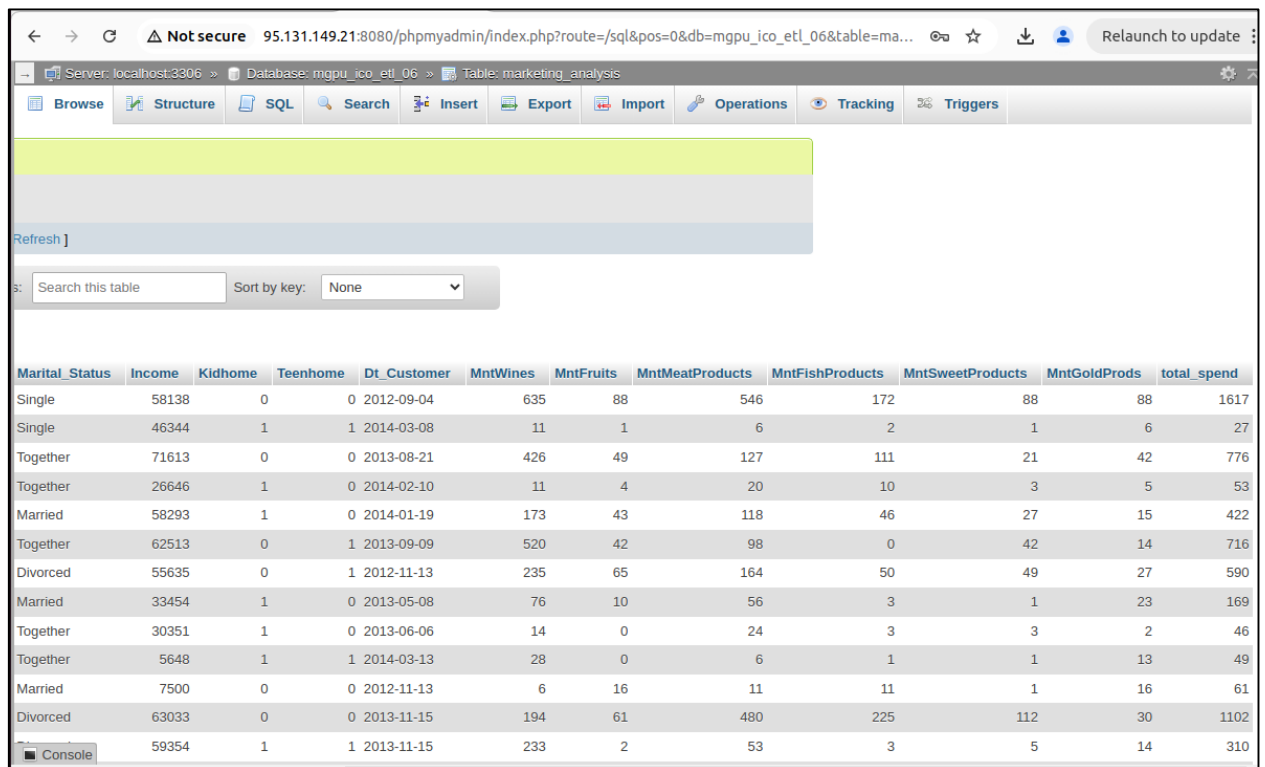
Extra options

	ID	Year_Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	MntWines	MntFruits	MntMeatProducts	MntFis
<input type="checkbox"/> Edit Copy Delete	1	1957	Graduation	Single	58138	0	0	2012-09-04	635	88	546	
<input type="checkbox"/> Edit Copy Delete	2	1954	Graduation	Single	46344	1	1	2014-03-08	11	1	6	
<input type="checkbox"/> Edit Copy Delete	3	1965	Graduation	Together	71613	0	0	2013-08-21	426	49	127	
<input type="checkbox"/> Edit Copy Delete	4	1984	Graduation	Together	26646	1	0	2014-02-10	11	4	20	
<input type="checkbox"/> Edit Copy Delete	5	1981	PhD	Married	58293	1	0	2014-01-19	173	43	118	
<input type="checkbox"/> Edit Copy Delete	6	1967	Master	Together	62513	0	1	2013-09-09	520	42	98	
<input type="checkbox"/> Edit Copy Delete	7	1971	Graduation	Divorced	55635	0	1	2012-11-13	235	65	164	
<input type="checkbox"/> Edit Copy Delete	8	1985	PhD	Married	33454	1	0	2013-05-08	76	10	56	
<input type="checkbox"/> Edit Copy Delete	9	1974	PhD	Together	30351	1	0	2013-06-06	14	0	24	
<input type="checkbox"/> Edit Copy Delete	10	1950	PhD	Together	5648	1	1	2014-03-13	28	0	6	
<input type="checkbox"/> Edit Copy Delete	11	1976	Basic	Married	7500	0	0	2012-11-13	6	16	11	
<input type="checkbox"/> Edit Copy Delete	12	1959	Graduation	Divorced	63033	0	0	2013-11-15	194	61	480	
<input type="checkbox"/> Edit Copy Delete	13	1952	Master	Divorced	59354	1	1	2013-11-15	233	2	53	

Console

Рис. 18 – Загруженные в таблицу данные

Также появился новый столбец, который был создан посредством калькулятора (Рис. 19).



The screenshot shows the phpMyAdmin interface for a database named 'mgpu_ico_etl_06'. The table 'marketing_analysis' is selected. The table structure is as follows:

Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	total_spend
Single	58138	0	0	2012-09-04	635	88	546	172	88	88	1617
Single	46344	1	1	2014-03-08	11	1	6	2	1	6	27
Together	71613	0	0	2013-08-21	426	49	127	111	21	42	776
Together	26646	1	0	2014-02-10	11	4	20	10	3	5	53
Married	58293	1	0	2014-01-19	173	43	118	46	27	15	422
Together	62513	0	1	2013-09-09	520	42	98	0	42	14	716
Divorced	55635	0	1	2012-11-13	235	65	164	50	49	27	590
Married	33454	1	0	2013-05-08	76	10	56	3	1	23	169
Together	30351	1	0	2013-06-06	14	0	24	3	3	2	46
Together	5648	1	1	2014-03-13	28	0	6	1	1	13	49
Married	7500	0	0	2012-11-13	6	16	11	11	1	16	61
Divorced	63033	0	0	2013-11-15	194	61	480	225	112	30	1102
	59354	1	1	2013-11-15	233	2	53	3	5	14	310

Рис. 19 – столбец рассчитанных в pentaho данных «total_spend»

Попробуем выполнить простой запрос на вывод записей по клиентам, сумма общих расходов которых больше 1000 (Рис. 20).

Server: localhost:3306 > Database: mgpu_ico_etl_06 > Table: marketing_analysis

Showing rows 0 - 24 (598 total, Query took 0.0013 seconds.) [total_spend: 1001... - 1027...]

SELECT * FROM `marketing_analysis` WHERE total_spend > 1000 ORDER BY `marketing_analysis`.`total_spend` ASC

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> Number of rows: 25 Filter rows: Search this table Sort by key: None

Extra options

	ID	Income	Dt_Customer	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	total_spend
<input type="checkbox"/> Edit Copy Delete	1672	58350	2013-01-04	493	26	206	116	80	80	1001
<input type="checkbox"/> Edit Copy Delete	2063	56796	2013-02-16	656	38	161	62	47	37	1001
<input type="checkbox"/> Edit Copy Delete	904	73113	2013-12-26	741	19	154	50	9	28	1001
<input type="checkbox"/> Edit Copy Delete	439	56796	2013-02-16	656	38	161	62	47	37	1001
<input type="checkbox"/> Edit Copy Delete	880	76630	2014-01-14	255	31	446	40	56	175	1003
<input type="checkbox"/> Edit Copy Delete	687	78468	2014-04-09	434	22	388	104	22	34	1004
<input type="checkbox"/> Edit Copy Delete	1151	80685	2012-08-22	241	45	604	34	26	54	1004
<input type="checkbox"/> Edit Copy Delete	375	63381	2012-10-05	571	50	142	33	50	159	1005
<input type="checkbox"/> Edit Copy Delete	628	63381	2012-10-05	571	50	142	33	50	159	1005
<input type="checkbox"/> Edit Copy Delete	743	63855	2013-02-09	359	35	314	93	116	89	1006
<input type="checkbox"/> Edit Copy Delete	417	61314	2013-04-25	378	0	189	97	172	172	1008
<input type="checkbox"/> Edit Copy Delete	542	81698	2013-11-06	179	28	520	111	123	47	1008
<input type="checkbox"/> Edit Copy Delete	231	64961	2012-12-22	282	114	276	75	124	28	1009

Рис. 20 – Результат запроса к столбцу «total_spend»

Вывод: в ходе выполнения работы было настроено подключение к БД, также изучен механизм работы трансформации в pentaho. По индивидуальному заданию были обработаны данные, а также рассчитаны новые значения на основе имеющегося датасета. Все данные удалось успешно загрузить в БД и выполнить запросы к таблице.