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

Макарова Екатерина Павловна

ЛАБОРАТОРНАЯ РАБОТА 1

Установка и настройка ETL-инструмента. Создание конвейеров данных

Проектный практикум по разработке ETL-решений

Направление подготовки

38.03.05 Бизнес-информатика

Профиль подготовки

Аналитика данных и эффективное управление

Курс обучения: 4

Форма обучения: очная

Преподаватель: кандидат технических наук,
доцент Босенко Тимур Муртазович

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.
- Проверка установки Java и WebKitGTK.
- 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 8: Анализ веб-аналитики: обработка данных о посещаемости сайта

Пошаговое выполнение:

1. Установка и запуск Pentaho Data Integration

Запуск Pentaho (Рис. 1).

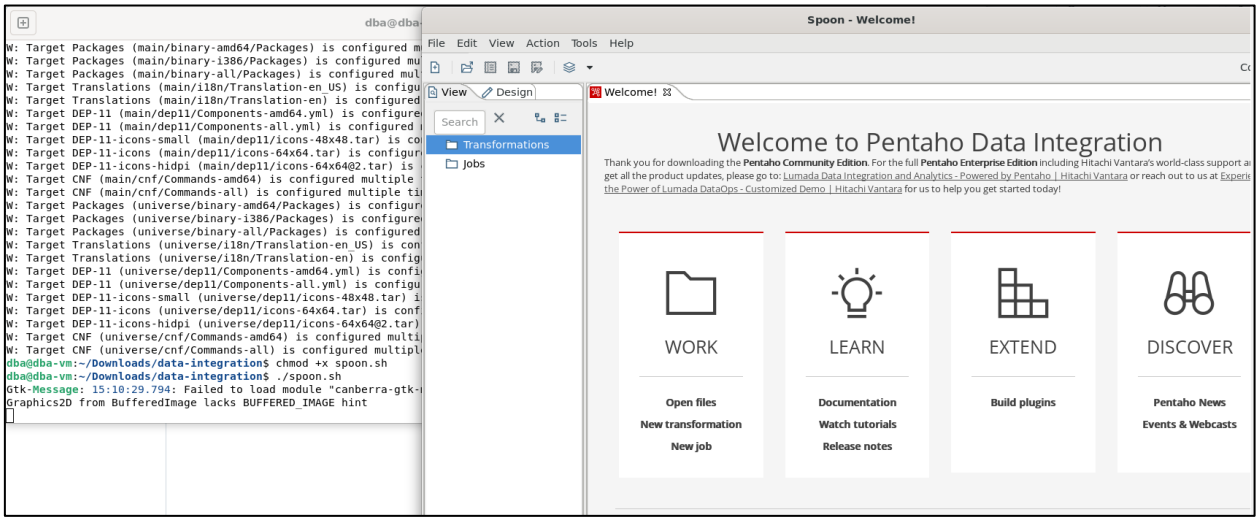


Рис. 1

2. **Установка MySQL драйвера для Pentaho Data Integration**

Установка драйвера MySQL (Рис. 2).

```
dba@dba-vm:~/Downloads$ sudo dpkg -i mysql-connector-j_9.2.0-1ubuntu22.04_all.deb
[sudo] password for dba:
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. 2

3. **Импорт данных**

Источник данных: https://www.kaggle.com/datasets/afranur/web-analytics-dataset

Добавление узла импорта файла CSV ().

Выбор скачанного CSV файла при добавлении в узел (Рис. 3).

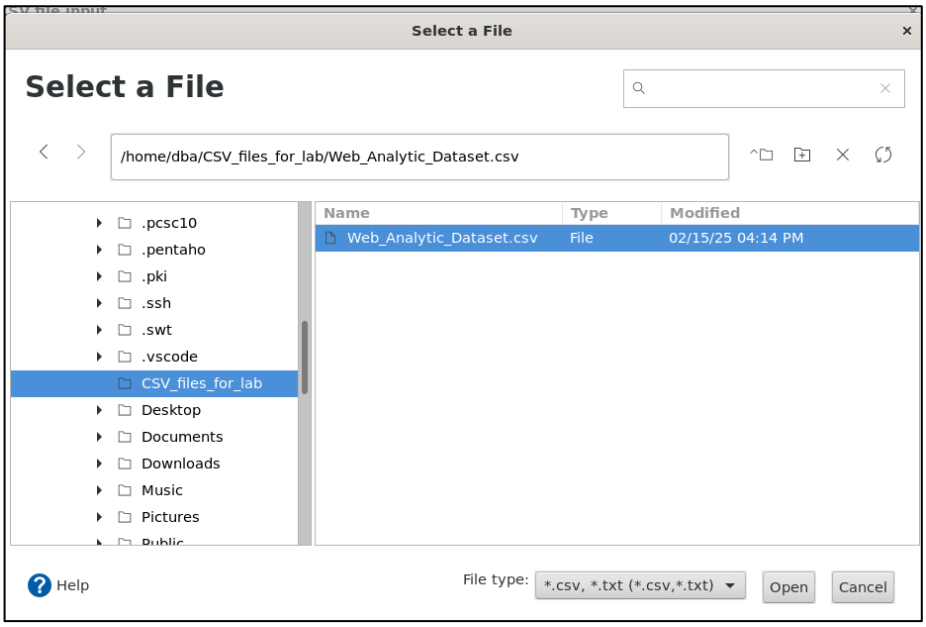


Рис. 3

Настройка узла импорта данных CSV (Рис. 4).

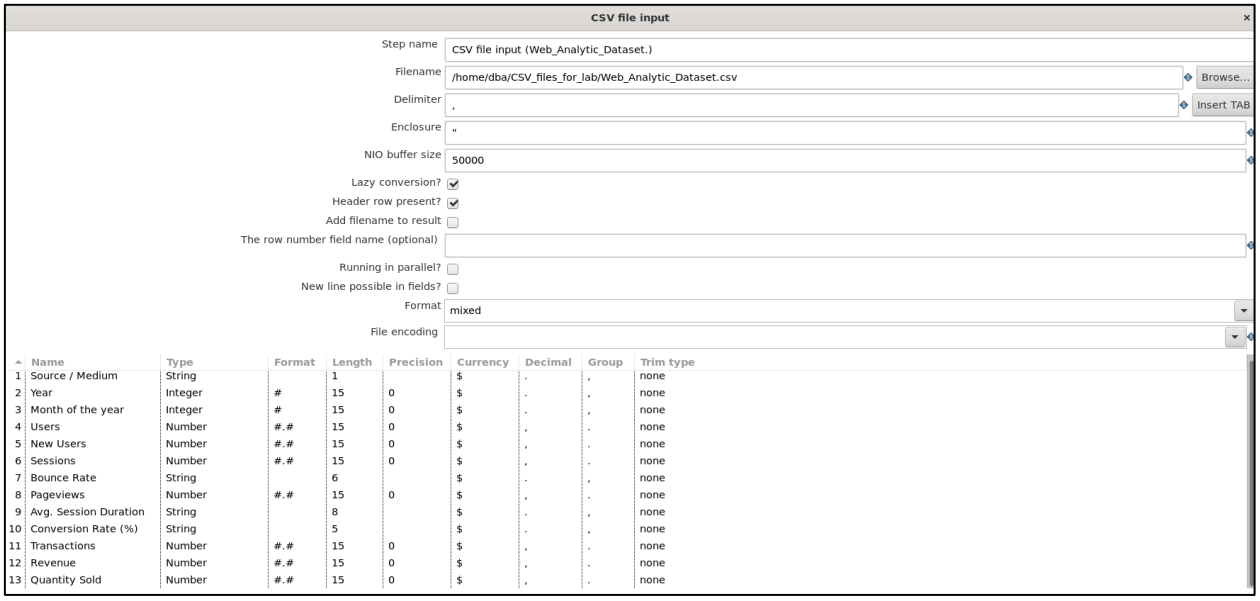


Рис. 4

В таблице 1 приведена информация о столбцах данных (наименование, тип и обозначение).

Таблица 1 – Тип данных в датасете

Наименование столбца	Тип данных	Значение
Source / Medium	String	Источник
Year	integer	Год

Month of the year	integer	Месяц
Users	integer	Количество пользователей
New Users	integer	Количество новых пользователей
Sessions	integer	Сеансы
Bounce Rate	string	Показатель отказов
Page views	integer	Просмотры страниц
Avg. Session Duration	string	Средняя продолжительность сеанса
Conversion Rate (%)	String	Коэффициент конверсии
Transactions	integer	Транзакции
Revenue	integer	Выручка
Quantity Sold	integer	Количество проданных товаров

При импорте данных их CSV типы данных не подходят для анализа:

1. Bounce Rate и Conversion Rate (%) являются строкой;
2. AVG. Session Duration в формате string, данные не нормализованы, необходимо привести например к секундам.

Предварительный просмотр полученных данных (Рис. 5).

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

First rows | Last rows | Off

	Source / Medium	Year	Month of the year	Users	New Users	Sessions	Bounce Rate	Pageviews	Avg. Session Duratio	Conversion Rate (%
1	A	2019	11	126.87	104.02	194.667	71.59%	455.159	00:01:11	0.2
2	A	2020	5	120.625	98.574	194.114	64.56%	559.509	00:01:32	0.69
3	A	2019	10	123.361	104.308	181.175	41.91%	368.907	00:01:05	0.26
4	A	2019	9	106.551	88.428	170.329	75.92%	368.803	00:01:01	0.18
5	A	2020	6	102.123	82.461	163.446	67.10%	425.41	00:01:20	0.7
6	A	2019	12	91.043	70.326	142.637	67.06%	370.798	00:01:20	0.34
7	A	2020	1	83.031	64.103	133.736	69.46%	373.356	00:01:23	0.45
8	A	2020	7	84.343	73.239	125.423	71.16%	292.263	00:01:12	0.58
9	A	2020	2	82.626	68.145	125.318	70.06%	328.822	00:01:24	0.52
10	A	2020	3	73.844	61.557	110.546	72.19%	266.187	00:01:15	0.48
11	B	2019	12	88.579	75.361	106.966	56.20%	333.48	00:01:55	0.21
12	A	2020	4	73.349	62.794	104.146	67.12%	288.195	00:01:25	0.6
13	C	2020	5	68.869	55.769	99.838	50.98%	325.311	00:01:38	0.47
14	B	2020	1	82.671	72.52	99.5	60.96%	299.723	00:01:48	0.25
15	B	2019	11	79.783	65.883	98.4	55.17%	342.257	00:02:13	0.22
16	D	2020	5	82.246	81.585	90.447	86.13%	111.745	00:00:26	0
17	B	2020	2	69.129	58.861	82.905	61.34%	238.736	00:01:51	0.35
18	B	2019	10	65.953	55.972	80.301	36.72%	263.957	00:02:14	0.33
19	C	2020	4	50.516	43.034	74.19	50.75%	247.484	00:01:35	0.38
20	B	2020	5	55.342	48.064	66.18	58.30%	213.302	00:02:07	0.52
21	E	2019	10	33.658	4.533	62.972	34.04%	164.136	00:01:46	0.25
22	B	2020	6	50.645	44.609	60.789	60.44%	181.511	00:01:58	0.53
23	B	2020	3	48.288	41.139	58.58	61.19%	174.115	00:02:00	0.31
24	A	2020	8	42.747	35.935	56.931	64.81%	149.505	00:01:29	0.48
25	E	2019	11	31.37	3.721	55.292	62.09%	143.387	00:01:42	0.16

Рис. 5

Bounce Rate содержит символы %, необходимо очистить данные от символов.

4. Обработка данных

4.1. Добавление узла Replace in string

Связываем импорт данных из CSV с новым узлом «Replace in string» (Рис. 6).

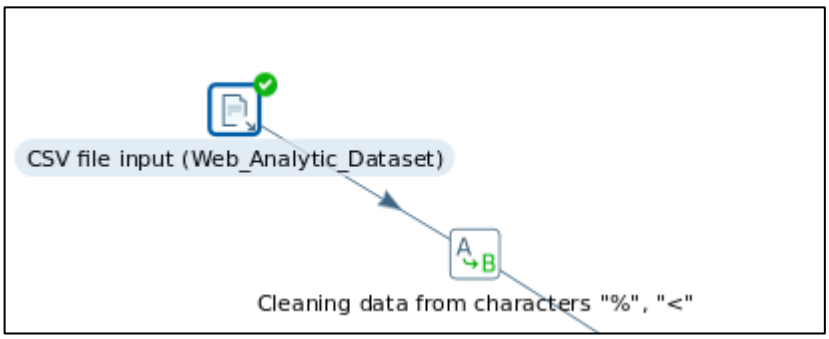


Рис. 6

Настройка узла «Replace in string»: очистка строковых данных от символов для дальнейшего изменения данных в числовой тип (Рис. 7).

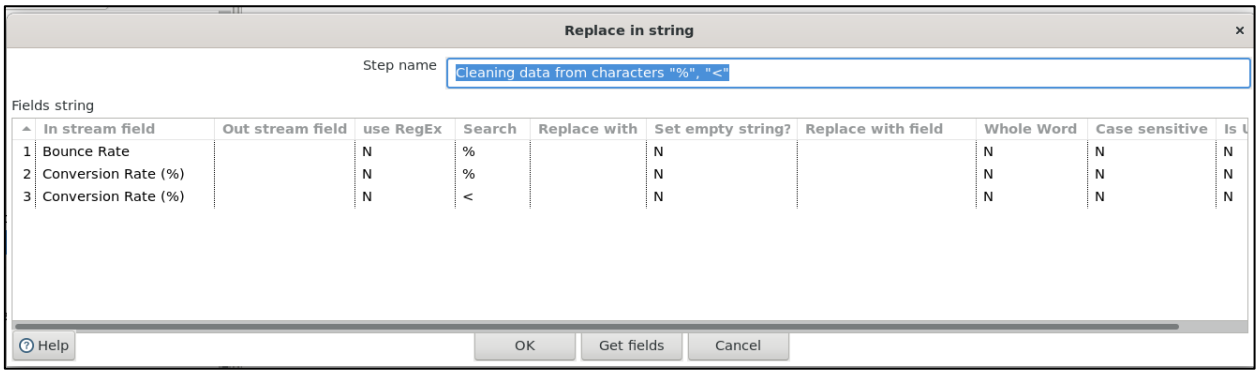


Рис. 7

4.2. Добавление компонента «Select values»

Изменение типов данных осуществляется с помощью компонента «Select values», связали предыдущей узел обработки данных с новым компонентом для изменения строковых данных в числовые (Рис. 8).

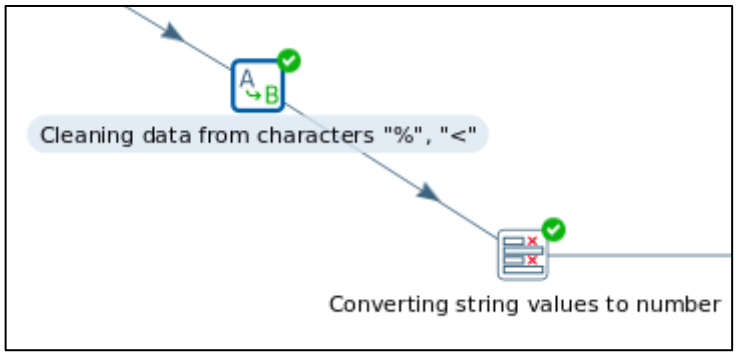


Рис. 8

Настройка узла для изменения столбцов «Bounce Rate» и «Conversion Rate (%)» с string на number (Рис. 9).

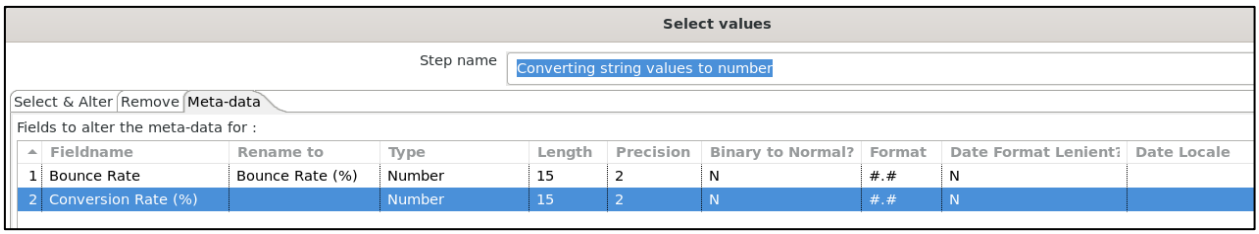


Рис. 9

4.3. Добавление компонента «Split fields»

Компонент «Split fields» разбивает столбец данных на несколько (Рис. 10).

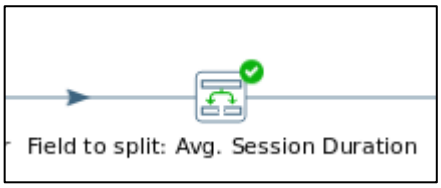


Рис. 10

Настройка узла «Split fields» для нормализации «Avg. Session Duration» (Рис. 11). Столбец, содержащий данные по средней продолжительности сеанса в формате часы: минуты: секунды, был разбит на три отдельных:

- Hours (часы);

- Minutes (минуты);

- Seconds (секунды).

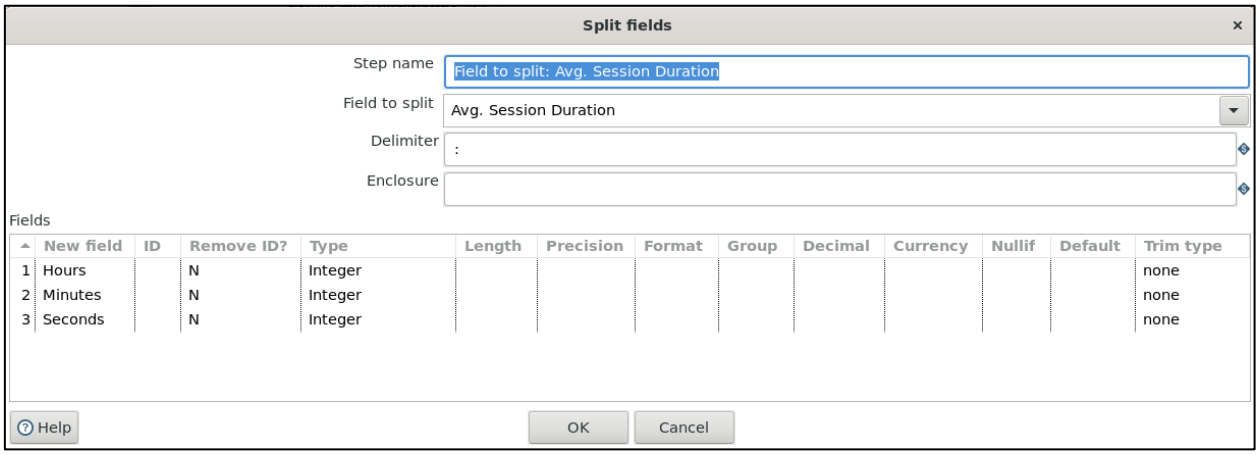


Рис. 11

Данная трансформация позволит привести к одному формату – общее количество секунд.

4.4. Добавление компонента «Modified JavaScript Value»

Далее полученные данные были преобразованы в общее количество секунд с помощью компонента с JavaScript (Рис. 12).

Рис. 12

Настройка узла «Modified JavaScript value»:

Создана новая переменная для расчёта общего количества секунд и записана в новый атрибут total_seconds (Рис. 13).

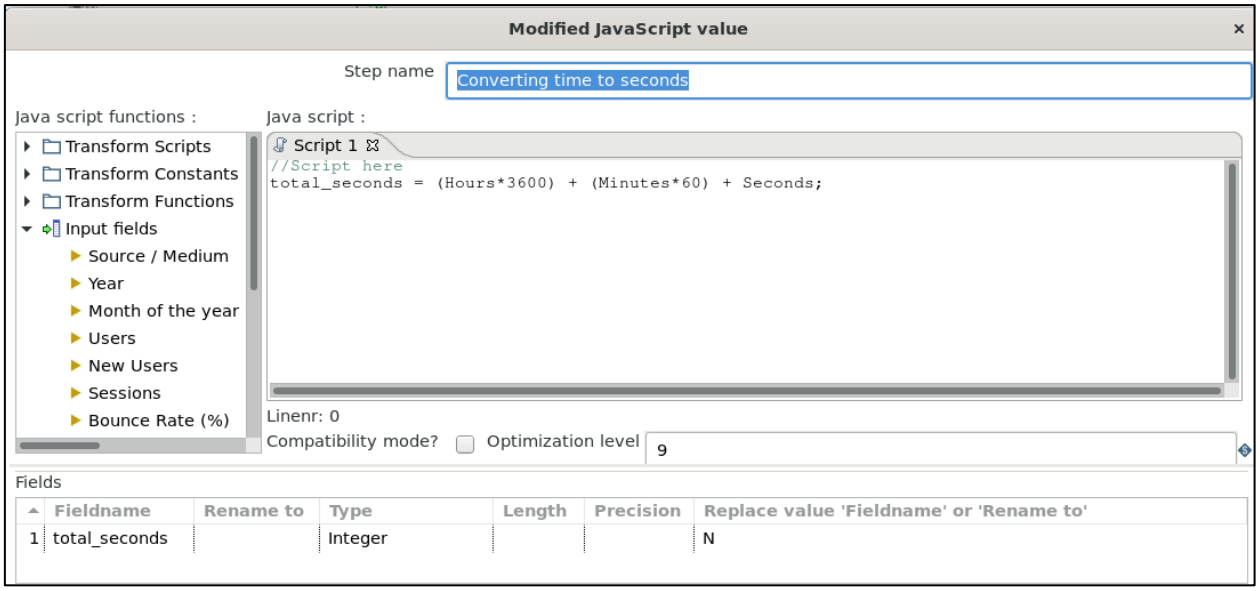


Рис. 13

4.5. Расчёт новой метрики «Revenue per Users/ Доход на пользователя»

Добавлен новый компонент «Calculator» для расчёта новой метрики (Рис. 14).

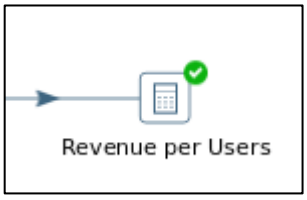


Рис. 14

Настройка узла «Calculator» для расчёта метрики «Доход на пользователя» (Рис. 15).

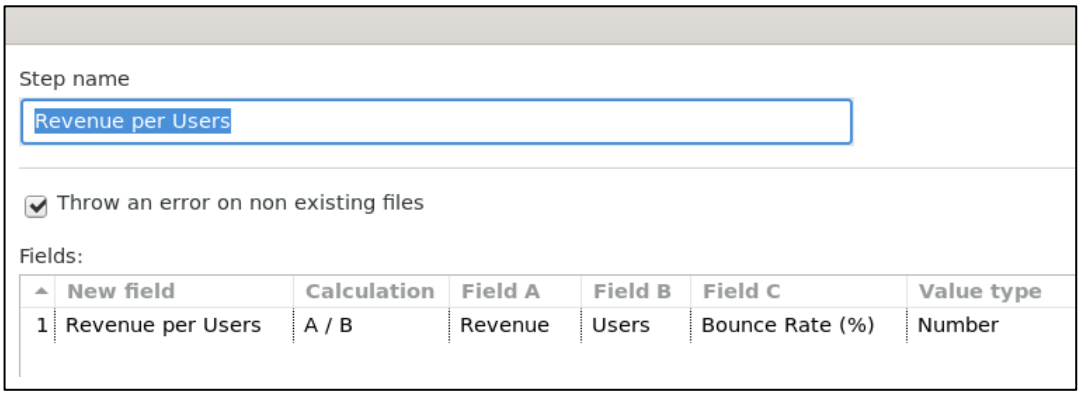


Рис. 15

Добавление узла «Table output» и его настройка:

1) Настройка Database Connection (Рис. 16).

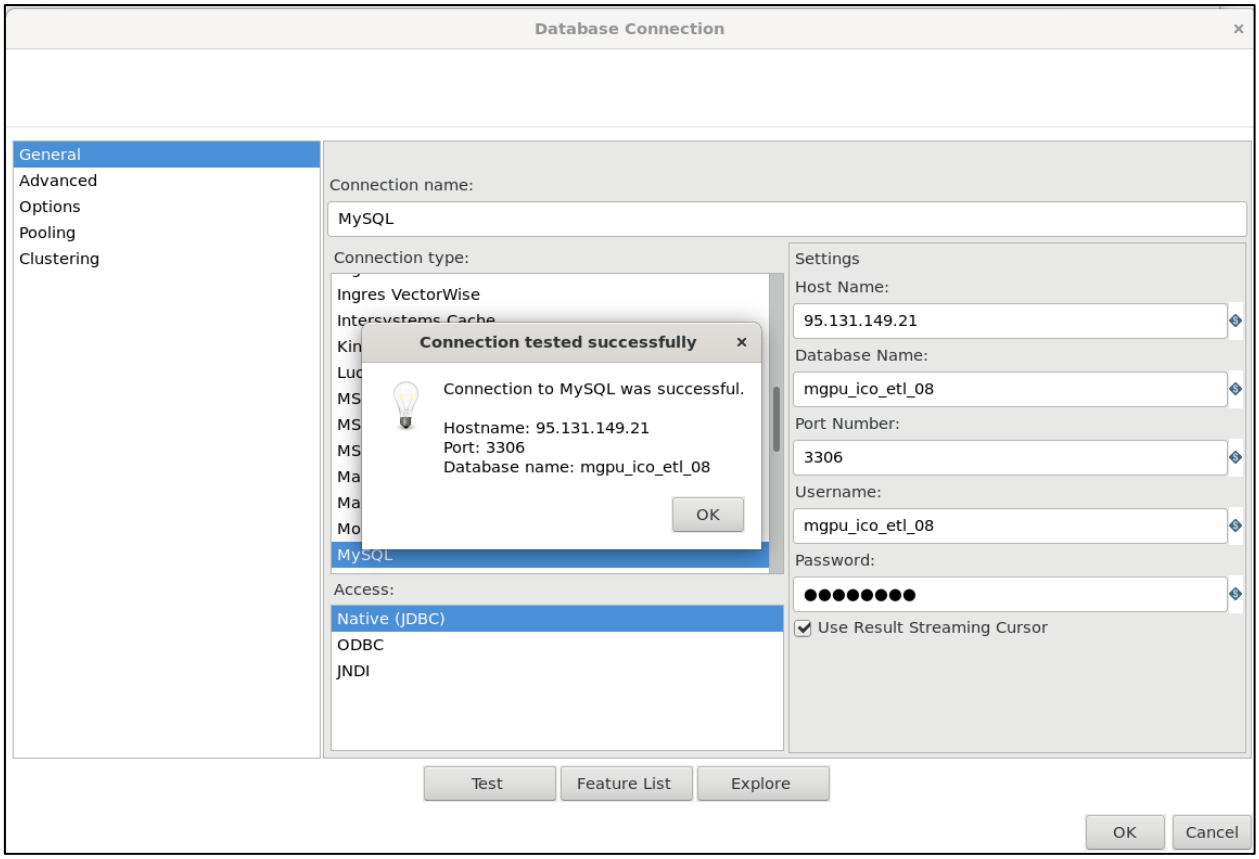


Рис. 16

2) Создание таблицы «web_analytic_data» с столбцами для последующего импорта данных в таблицу (Рис. 17).

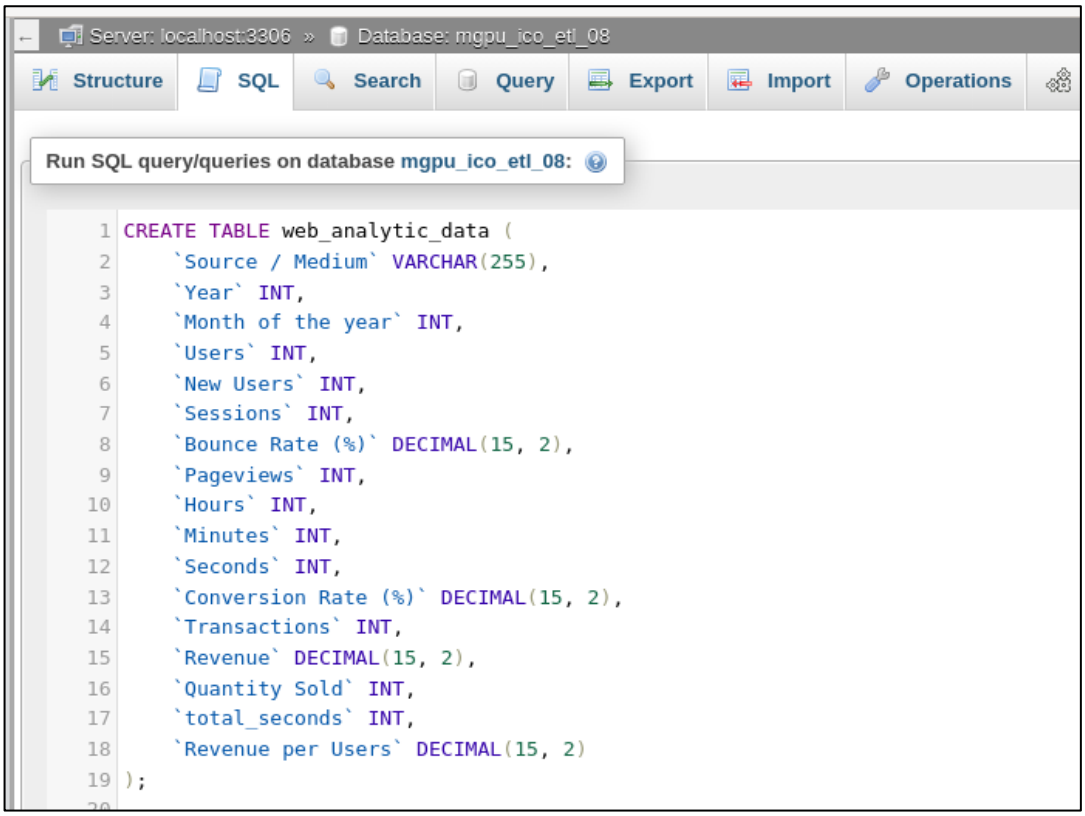


Рис. 17

3) Настройка узла Table output: указали shema и table для импорта данных (Рис. 18).

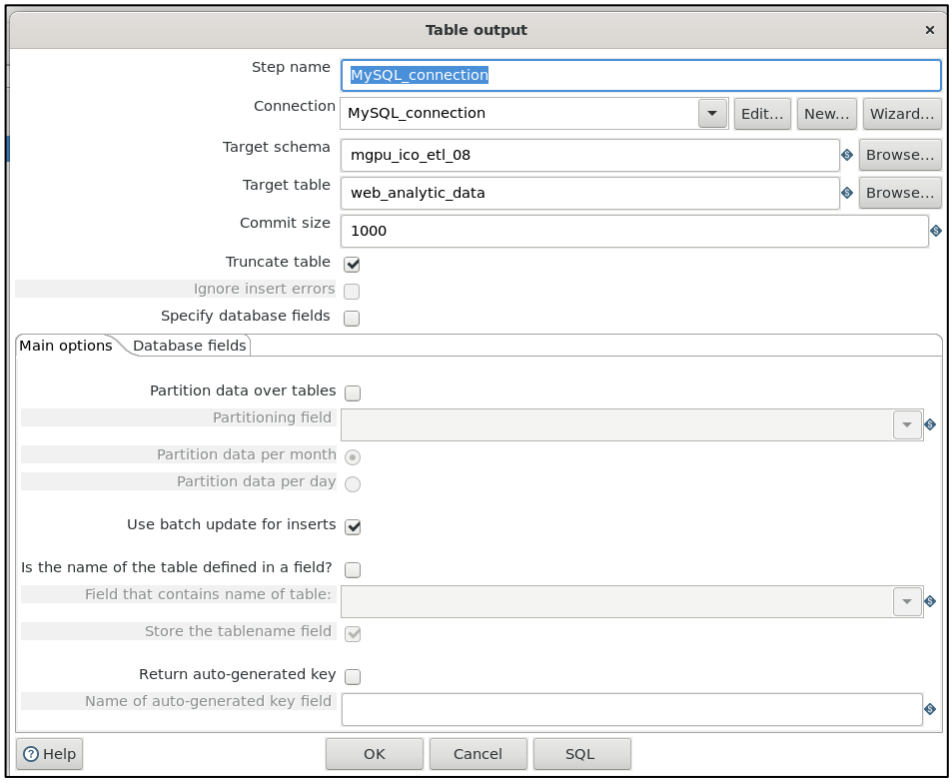


Рис. 18

Полный вид трансформации (Рис. 19).

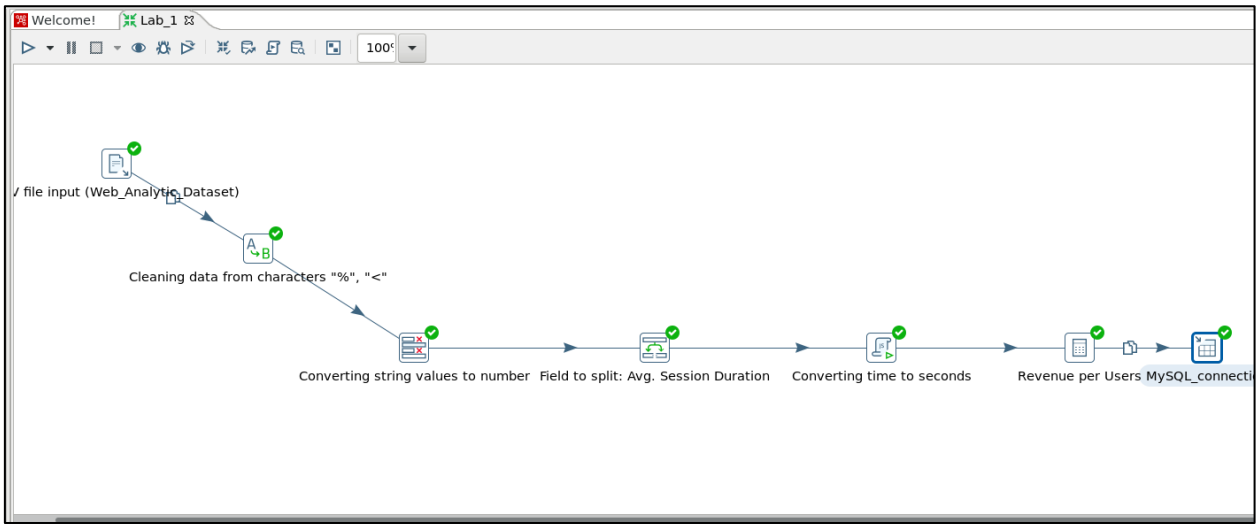
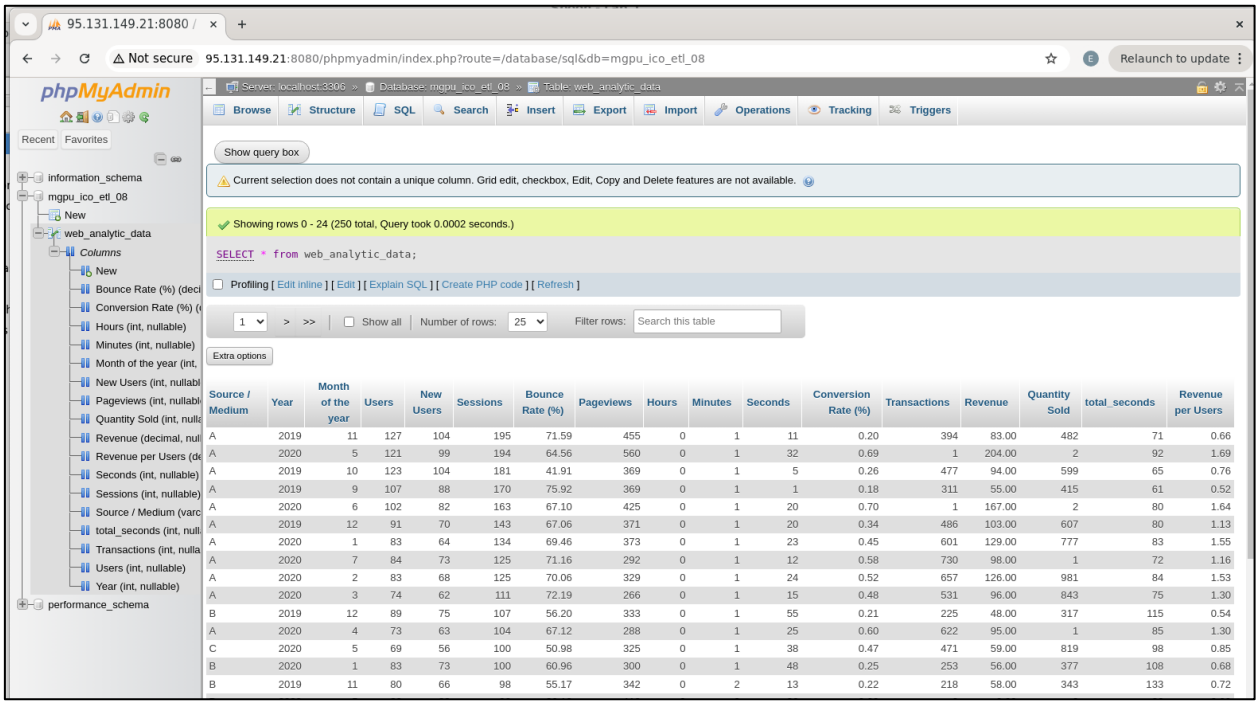


Рис. 19

SQL запрос к таблице на проверку импорта данных (Рис. 20).



Source / Medium	Year	Month of the year	Users	New Users	Sessions	Bounce Rate (%)	Pageviews	Hours	Minutes	Seconds	Conversion Rate (%)	Transactions	Revenue	Quantity Sold	total_seconds	Revenue per Users
A	2019	11	127	104	195	71.59	455	0	1	11	0.20	394	83.00	482	71	0.66
A	2020	5	121	99	194	64.56	560	0	1	32	0.69	1	204.00	2	92	1.69
A	2019	10	123	104	181	41.91	369	0	1	5	0.26	477	94.00	599	65	0.76
A	2019	9	107	88	170	75.92	369	0	1	1	0.18	311	55.00	415	61	0.52
A	2020	6	102	82	163	67.10	425	0	1	20	0.70	1	167.00	2	80	1.64
A	2019	12	91	70	143	67.06	371	0	1	20	0.34	486	103.00	607	80	1.13
A	2020	1	83	64	134	69.46	373	0	1	23	0.45	601	129.00	777	83	1.55
A	2020	7	84	73	125	71.16	292	0	1	12	0.58	730	98.00	1	72	1.16
A	2020	2	83	68	125	70.06	329	0	1	24	0.52	657	126.00	981	84	1.53
A	2020	3	74	62	111	72.19	266	0	1	15	0.48	531	96.00	843	75	1.30
B	2019	12	89	75	107	56.20	333	0	1	55	0.21	225	48.00	317	115	0.54
A	2020	4	73	63	104	67.12	288	0	1	25	0.60	622	95.00	1	85	1.30
C	2020	5	69	56	100	50.98	325	0	1	38	0.47	471	59.00	819	98	0.85
B	2020	1	83	73	100	60.96	300	0	1	48	0.25	253	56.00	377	108	0.68
B	2019	11	80	66	98	55.17	342	0	2	13	0.22	218	58.00	343	133	0.72

Рис. 20

SQL запрос на агрегирование данных, который подсчитывает общее количество пользователей, среднее значение сессий и общую сумму доходов, сгруппированных по году (Рис. 21).

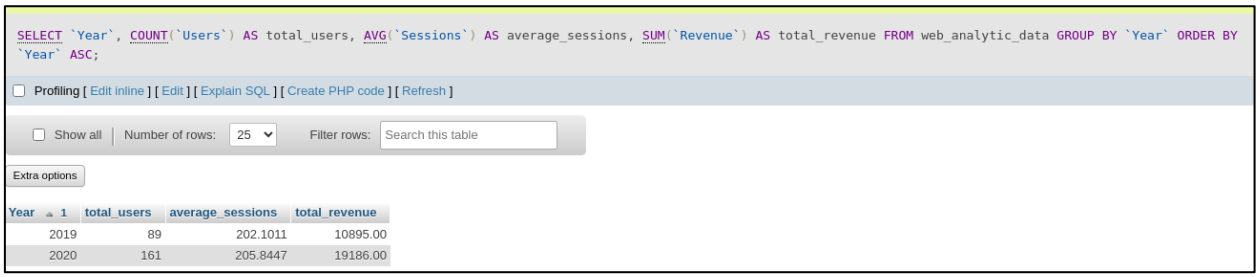

SELECT `Year`, COUNT(`Users`) AS total_users, AVG(`Sessions`) AS average_sessions, SUM(`Revenue`) AS total_revenue FROM web_analytic_data GROUP BY `Year` ORDER BY `Year` ASC;
Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]
Show all | Number of rows: 25 Filter rows: Search this table
Extra options
Year 1 total_users average_sessions total_revenue
2019 89 202.1011 10895.00
2020 161 205.8447 19186.00

Рис. 21