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

Установка и настройка ETL-инструмента. Создание конвейеров данных

Выполнила: Ванярина Ю.А., группа: АДЭУ-211
Преподаватель: Босенко Т.М.

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

ХОД РАБОТЫ:

Выполнено подключение к виртуальной машине и запуск Pentaho (рисунок 1).

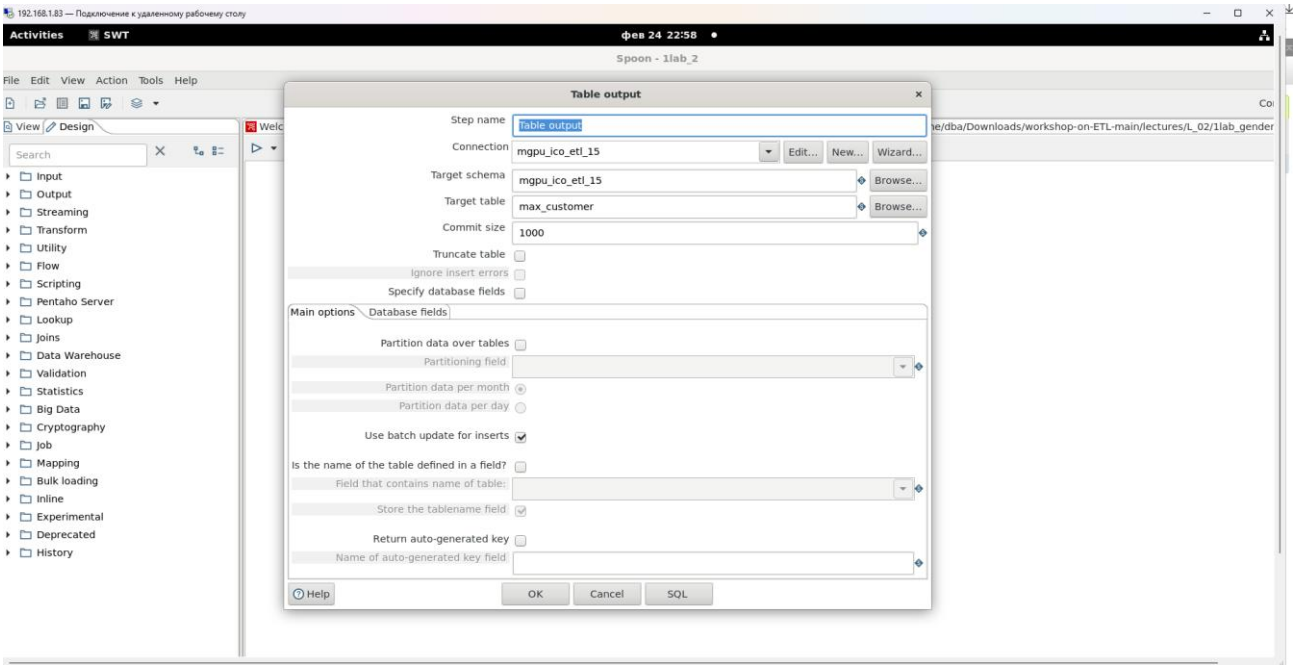


Рисунок 1 – настройка подключения к бд в Pentaho

Вариант 15.

Аналитика телекоммуникаций: анализ данных об использовании услуг Telecom Dataset (https://www.kaggle.com/datasets/blastchar/telco-customer-churn)

1. Датасет был скачен с kaggle и изучен (таблица 1)

Таблица 1 – описание столбцов датасета

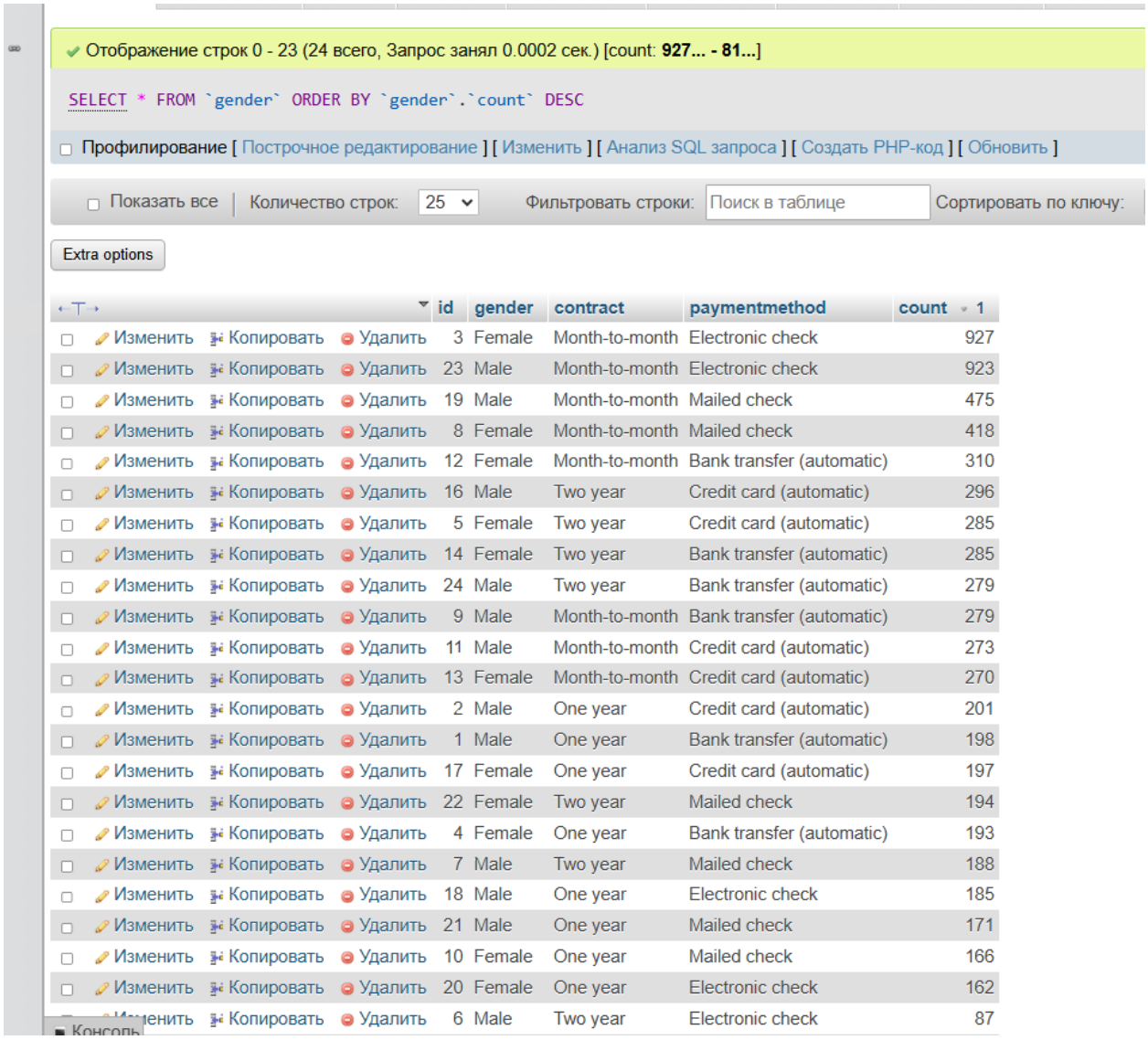
customerID	Уникальный идентификатор клиента
gender	Пол клиента (Male, Female)
SeniorCitizen	Является ли клиент пожилым (1 - Да, 0 - Нет)
Partner	Есть ли у клиента супруг/супруга (Yes, No)
Dependents	Есть ли у клиента иждивенцы (Yes, No)
tenure	Сколько месяцев клиент пользуется услугами
PhoneService	Подключена ли телефонная связь (Yes, No)
MultipleLines	Подключены ли несколько линий (Yes, No, No phone service)
InternetService	Тип интернет-соединения (DSL, Fiber optic, No)
OnlineSecurity	Подключена ли онлайн-защита (Yes, No, No internet service)

OnlineBackup	Подключено ли онлайн-резервное копирование (Yes, No, No internet service)
DeviceProtection	Подключена ли защита устройства (Yes, No, No internet service)
TechSupport	Есть ли техническая поддержка (Yes, No, No internet service)
StreamingTV	Подключено ли потоковое ТВ (Yes, No, No internet service)
StreamingMovies	Подключены ли потоковые фильмы (Yes, No, No internet service)
Contract	Тип контракта (Month-to-month, One year, Two year)
PaperlessBilling	Использует ли клиент электронный биллинг (Yes, No)
PaymentMethod	Способ оплаты (Electronic check, Mailed check, Bank transfer, Credit card)
MonthlyCharges	Сколько клиент платит в месяц ($)
TotalCharges	Сколько клиент заплатил за все время ($)
Churn	Целевая переменная: Ушел клиент (Yes - Да, No - Нет)

Для анализа составлено несколько отчётов, которые помогают сделать выводы о датасете.

1. Отчёт по гендеру, контракту и способу оплаты (рисунок 2)

На данном отчёте видна частота оплаты, например, женщины с контрактом типа ‘month-to-month’ чаще всего оплачивают контракт электронным чеком. Также этот тип оплаты и контракта является самым популярным и среди мужчин.



✔ Отображение строк 0 - 23 (24 всего, Запрос занял 0.0002 сек.) [count: **927... - 81...**]

```
SELECT * FROM `gender` ORDER BY `gender`.`count` DESC
```

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

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

Extra options

	id	gender	contract	paymentmethod	count 1
Изменить Копировать Удалить	3	Female	Month-to-month	Electronic check	927
Изменить Копировать Удалить	23	Male	Month-to-month	Electronic check	923
Изменить Копировать Удалить	19	Male	Month-to-month	Mailed check	475
Изменить Копировать Удалить	8	Female	Month-to-month	Mailed check	418
Изменить Копировать Удалить	12	Female	Month-to-month	Bank transfer (automatic)	310
Изменить Копировать Удалить	16	Male	Two year	Credit card (automatic)	296
Изменить Копировать Удалить	5	Female	Two year	Credit card (automatic)	285
Изменить Копировать Удалить	14	Female	Two year	Bank transfer (automatic)	285
Изменить Копировать Удалить	24	Male	Two year	Bank transfer (automatic)	279
Изменить Копировать Удалить	9	Male	Month-to-month	Bank transfer (automatic)	279
Изменить Копировать Удалить	11	Male	Month-to-month	Credit card (automatic)	273
Изменить Копировать Удалить	13	Female	Month-to-month	Credit card (automatic)	270
Изменить Копировать Удалить	2	Male	One year	Credit card (automatic)	201
Изменить Копировать Удалить	1	Male	One year	Bank transfer (automatic)	198
Изменить Копировать Удалить	17	Female	One year	Credit card (automatic)	197
Изменить Копировать Удалить	22	Female	Two year	Mailed check	194
Изменить Копировать Удалить	4	Female	One year	Bank transfer (automatic)	193
Изменить Копировать Удалить	7	Male	Two year	Mailed check	188
Изменить Копировать Удалить	18	Male	One year	Electronic check	185
Изменить Копировать Удалить	21	Male	One year	Mailed check	171
Изменить Копировать Удалить	10	Female	One year	Mailed check	166
Изменить Копировать Удалить	20	Female	One year	Electronic check	162
Изменить Копировать Удалить	6	Male	Two year	Electronic check	87

Консоль

Рисунок 2 – отчёт по гендеру

Данный отчет был реализован с помощью узла memory group by в pentaho (рисунок 3)

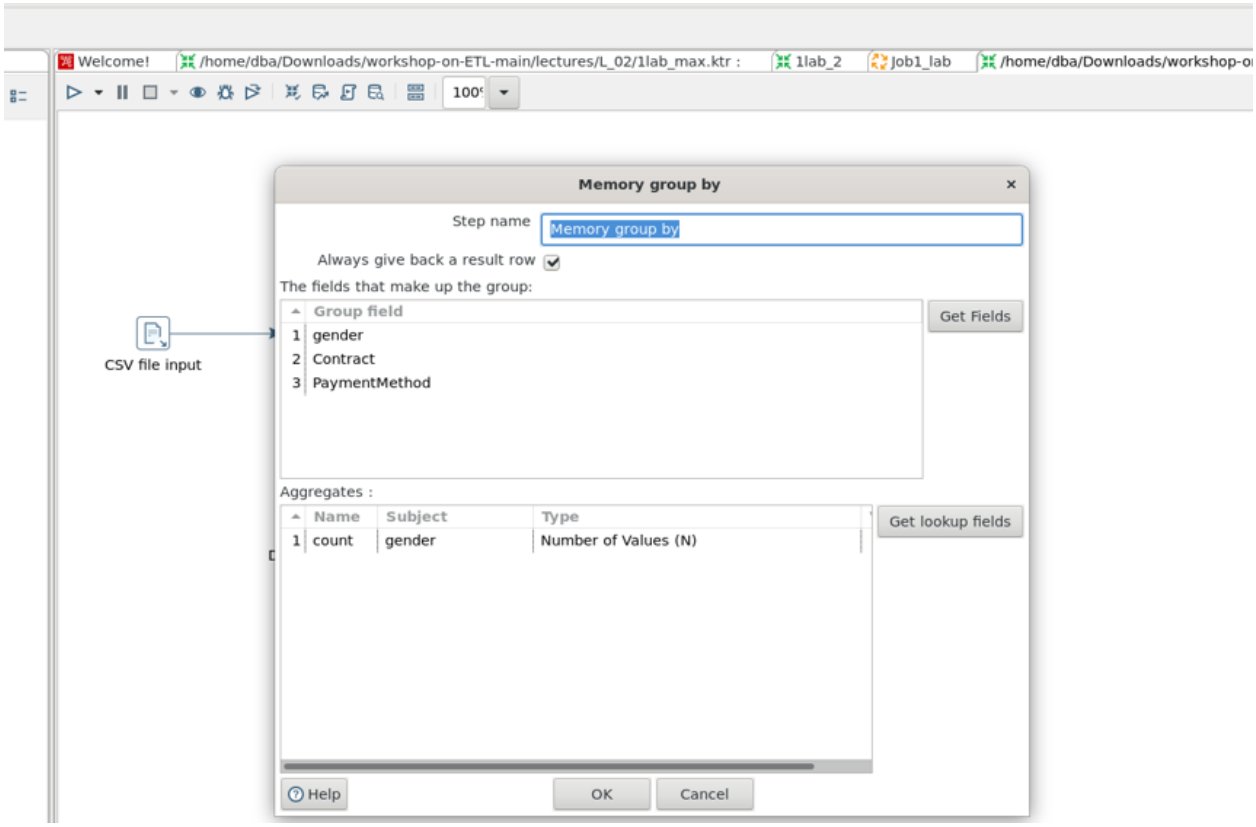


Рисунок 3 – настройка узла memory group by

Трансформация по первому отчёту выглядит следующим образом (рисунок 4):

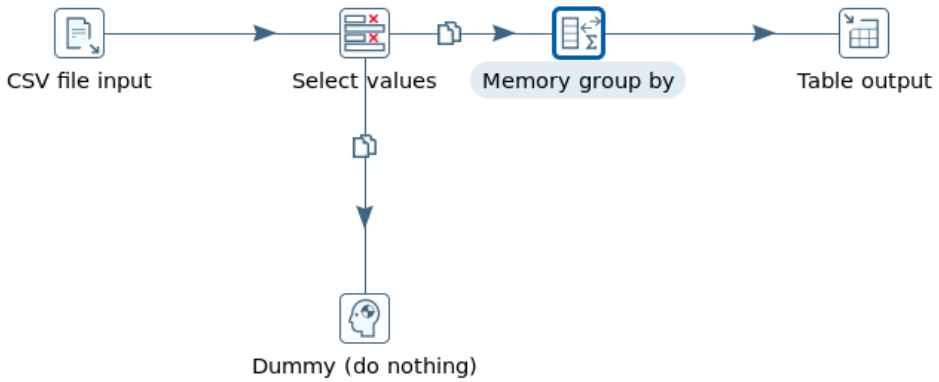


Рисунок 4 – трансформация 1

Отчёт был составлен с помощью логики следующего sql запроса:

```
select
e.gender ,
e.contract ,
e.paymentmethod ,
count(e.gender)
from etl_lab_1 e
group by 1,2,3
order by 1,2,3
```

2. Второй отчёт находит 1 пользователя, который больше всего заплатил totalcharges

Составление отчёта было реализовано, опираясь на следующий sql запроса:

```
select
e.customerid,e.gender,e.totalcharges,e.paymentmethod ,e.contract
from etl_lab_1 e
where
e.totalcharges = (select max(e1.totalcharges) from etl_lab_1 e1 )
```

такой запрос был разбит на 2 трансформации

Первая создавала отчёт по запросу (рисунок 5.1):

```
select
e.customerid,e.gender,e.totalcharges,e.paymentmethod ,e.contract
from etl_lab_1 e
```

данные были обработаны с помощью узла if field value is null (рисунок 5)

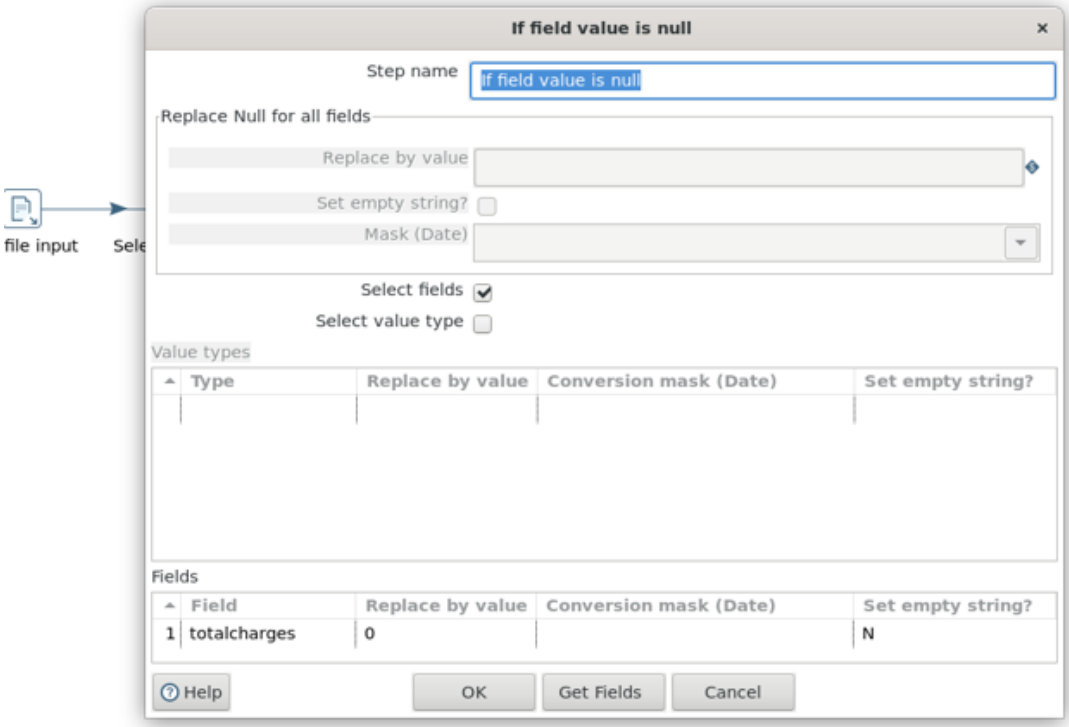


Рисунок 5 – настройка узла

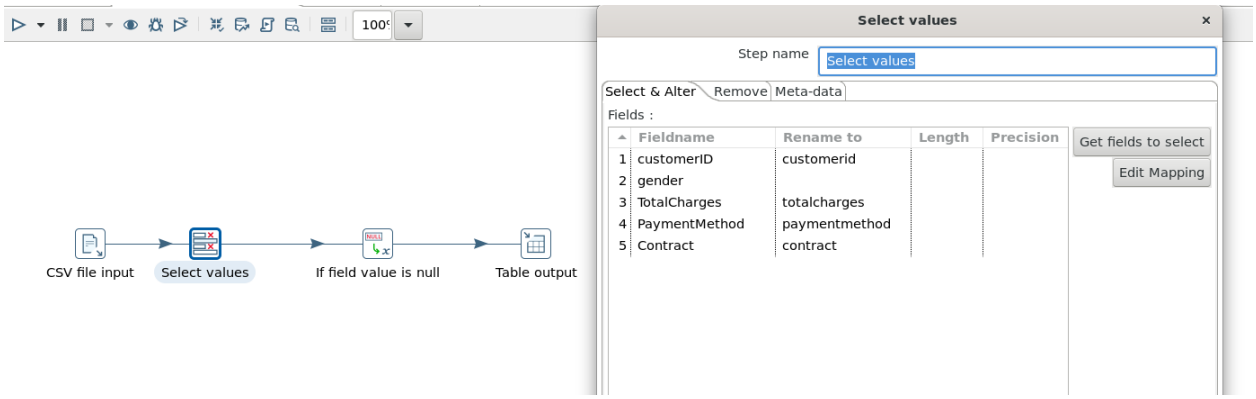


Рисунок 5.1 – первая трансформация

Вторая трансформация уже работала по конкретному запросу,где брала с базы данных созданную таблицу и с помощью sql выводила максимальное значение (рисунок 6)

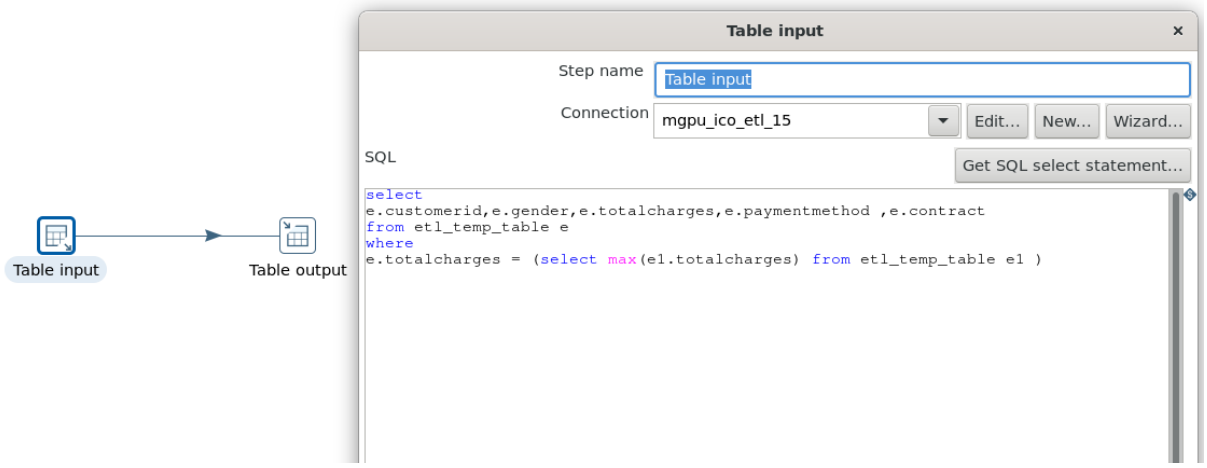


Рисунок 6 – вторая трансформация

В результате выполнения трансформаций был получен следующий результат (рисунок 7):

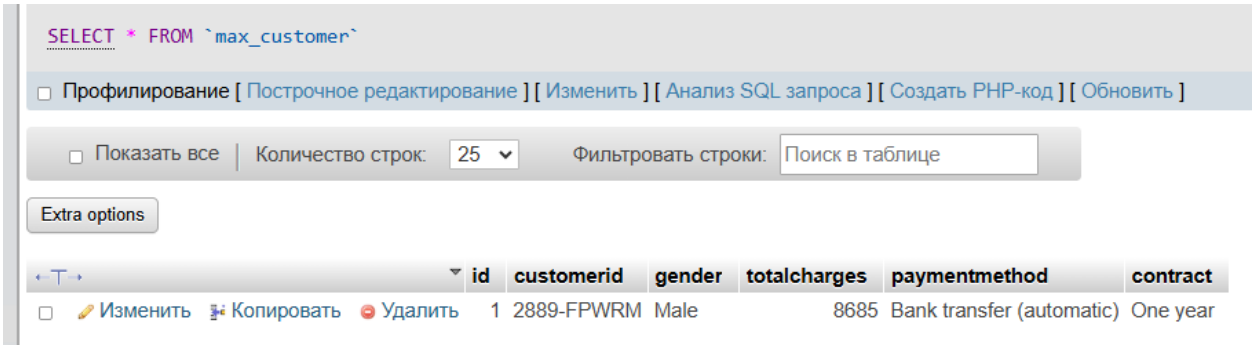


Рисунок 7 – результат трансформации

Отчёт показывает,что самый дорогой контракт был у мужчины с id 2889-FPWRM, который оплачивался автоматически и был длительностью на один год.

Для удобства все три трансформации были объединены в один job для удобного выполнения (рисунок 8)

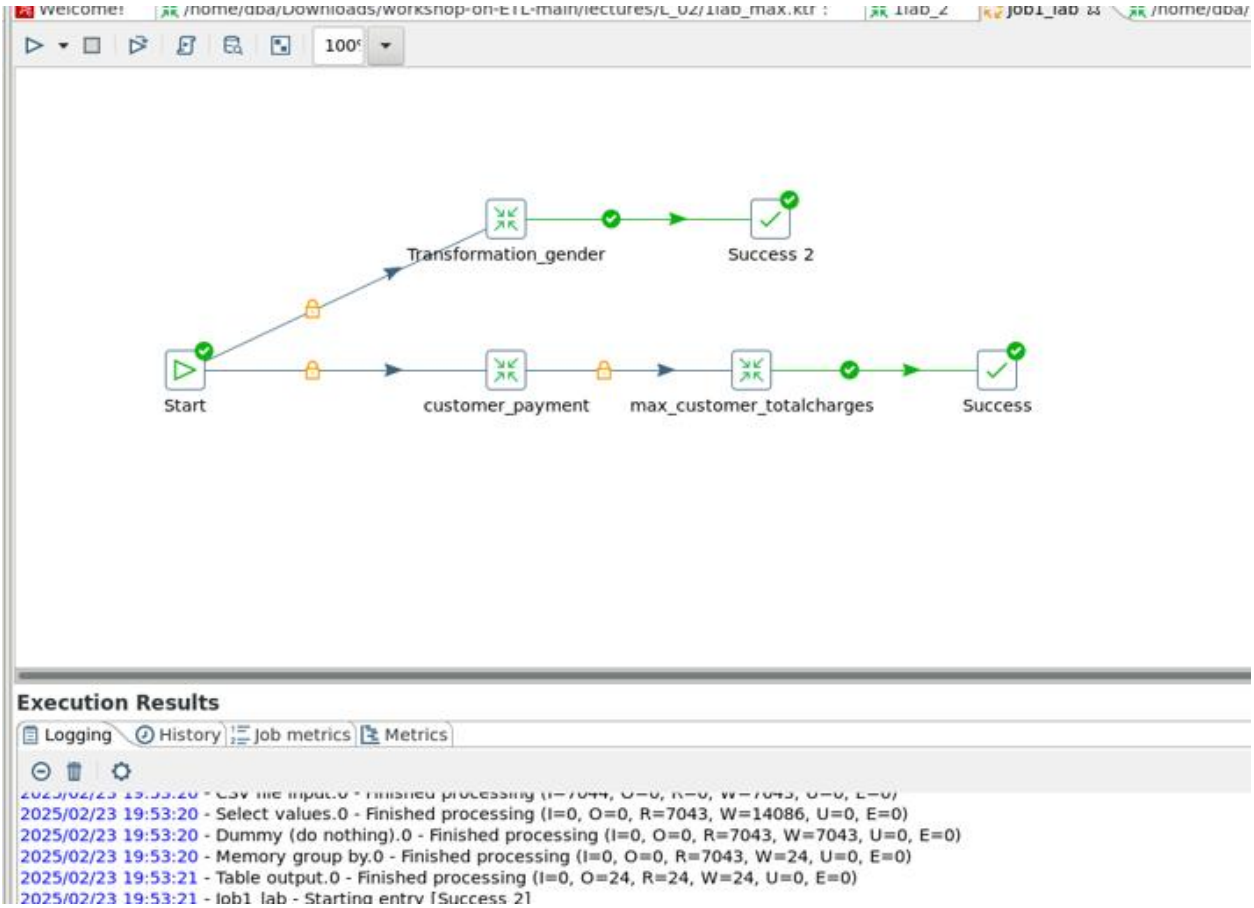


Рисунок 8 –job

Проверка через sql:

```sql
select
e.gender ,
e.contract ,
e.paymentmethod ,
count(e.gender)
from etl_lab_1 e
group by 1,2,3
order by 4 desc, 1,2,3
```

etl_lab_1 1 | etl_lab_1 2 | etl_lab_1 3 | etl_lab_1 4 | etl_lab_1 5

select e.gender , e.contract , e.paymentmethod , count(e.gender) from ... Введите SQL выражение чтобы ото

	gender	contract	paymentmethod	count
1	Female	Month-to-month	Electronic check	927
2	Male	Month-to-month	Electronic check	923
3	Male	Month-to-month	Mailed check	475
4	Female	Month-to-month	Mailed check	418
5	Female	Month-to-month	Bank transfer (automatic)	310
6	Male	Two year	Credit card (automatic)	296
7	Female	Two year	Bank transfer (automatic)	285
8	Female	Two year	Credit card (automatic)	285
9	Male	Month-to-month	Bank transfer (automatic)	279
10	Male	Two year	Bank transfer (automatic)	279
11	Male	Month-to-month	Credit card (automatic)	273
12	Female	Month-to-month	Credit card (automatic)	270
13	Male	One year	Credit card (automatic)	201
14	Male	One year	Bank transfer (automatic)	198
15	Female	One year	Credit card (automatic)	197
16	Female	Two year	Mailed check	194
17	Female	One year	Bank transfer (automatic)	193
18	Male	Two year	Mailed check	188
19	Male	One year	Electronic check	185
20	Male	One year	Mailed check	171
21	Female	One year	Mailed check	166
22	Female	One year	Electronic check	162
23	Male	Two year	Electronic check	87
24	Female	Two year	Electronic check	81

```
select
e.customerid,e.gender,e.totalcharges,e.paymentmethod ,e.contract
from etl_lab_1 e
```

etl_lab_1 1 | etl_lab_1 2 | etl_lab_1 3 | etl_lab_1 4 | etl_lab_1 5

select e.customerid,e.gender,e.totalcharges,e.paymentmethod ,e.contra — Введите SQL выражение чтобы отфильтровать результат

	customerid	gender	totalcharges	paymentmethod	contract
1	7590-VHVEG	Female	29,85	Electronic check	Month-to-month
2	5575-GNVDE	Male	1 889,5	Mailed check	One year
3	3668-QPYBK	Male	108,15	Mailed check	Month-to-month
4	7795-CFOCW	Male	1 840,75	Bank transfer (automatic)	One year
5	9237-HQITU	Female	151,65	Electronic check	Month-to-month
6	9305-CDSKC	Female	820,5	Electronic check	Month-to-month
7	1452-KIOVK	Male	1 949,4	Credit card (automatic)	Month-to-month
8	6713-OKOMC	Female	301,9	Mailed check	Month-to-month
9	7892-POOKP	Female	3 046,05	Electronic check	Month-to-month
10	6388-TABGU	Male	3 487,95	Bank transfer (automatic)	One year
11	9763-GRSKD	Male	587,45	Mailed check	Month-to-month
12	7469-LKBCI	Male	326,8	Credit card (automatic)	Two year
13	8091-TTVAX	Male	5 681,1	Credit card (automatic)	One year
14	0280-XJGEX	Male	5 036,3	Bank transfer (automatic)	Month-to-month
15	5129-JLPIS	Male	2 686,05	Electronic check	Month-to-month
16	3655-SNQYZ	Female	7 895,15	Credit card (automatic)	Two year
17	8191-XWSZG	Female	1 022,95	Mailed check	One year
18	9959-WOFKT	Male	7 382,25	Bank transfer (automatic)	Two year
19	4190-MFLUW	Female	528,35	Credit card (automatic)	Month-to-month
20	4183-MYFRB	Female	1 862,9	Electronic check	Month-to-month
21	8779-QRDMV	Male	39,65	Electronic check	Month-to-month
22	1680-VDCWW	Male	202,25	Bank transfer (automatic)	One year
23	1066-JKSGK	Male	20,15	Mailed check	Month-to-month
24	3638-WEABW	Female	3 505,1	Credit card (automatic)	Two year
25	6322-HRPFA	Male	2 970,3	Credit card (automatic)	Month-to-month
26	6865-JZNKO	Female	1 530,6	Bank transfer (automatic)	Month-to-month
27	6467-CHFZW	Male	4 749,15	Electronic check	Month-to-month
28	8665-UTDHZ	Male	30,2	Electronic check	Month-to-month
29	5248-YGIJN	Male	6 369,45	Credit card (automatic)	Two year
30	8773-HHUOZ	Female	1 093,1	Mailed check	Month-to-month
31	3841-NFECX	Female	6 766,95	Credit card (automatic)	Two year

```
select
e.customerid,e.gender,e.totalcharges,e.paymentmethod ,e.contract
from etl_lab_1 e
where
e.totalcharges = (select max(e1.totalcharges) from etl_lab_1 e1 )
```

etl_lab_1 1 | etl_lab_1 2 | etl_lab_1 3 | etl_lab_1 4 | etl_lab_1 5

select e.customerid,e.gender,e.totalcharges,e.paymentmethod ,e.contra — Введите SQL выражение чтобы отфильтровать

	customerid	gender	totalcharges	paymentmethod	contract
1	2889-FPWRM	Male	8 684,8	Bank transfer (automatic)	One year

Выводы: В ходе работы был успешно запущен pentaho и выполнено 3 трансформации для анализа данных, изучены группировка и работа с уже

существующими таблицами в базе данных. Построен job для более быстрой отработки трансформаций.