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

Нургалеева Гузель Рустэмовна БД-241м

Data Engineering. Практическая работа 4-3.
Вариант задания: 18

Направление подготовки/специальность
38.04.05 - Бизнес-информатика
Бизнес-аналитика и большие данные
(очная форма обучения)

Москва
2024

Цель работы: получить практические навыки интеграции, обработки и согласования данных из различных источников.

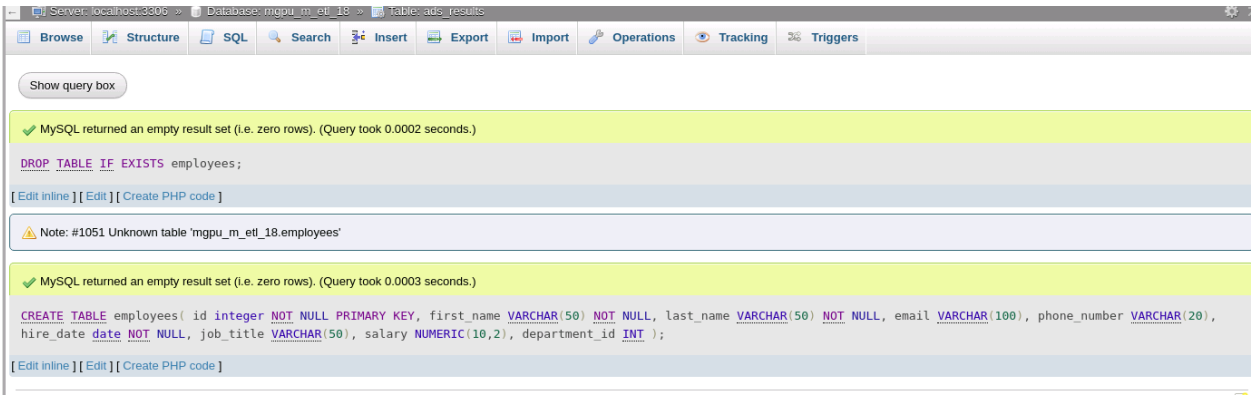
Задачи:
- Изучить методы чтения данных из разных источников.
- Освоить техники обработки и очистки данных.
- Научиться согласовывать данные из разных источников.
- Реализовать сохранение обработанных данных

Ход работы:

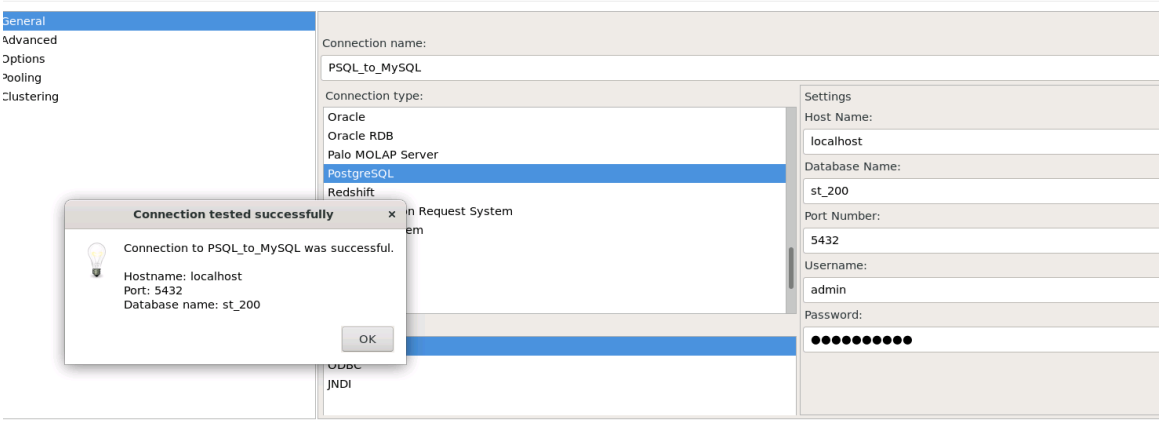
Создание таблиц в БД, в которые будут загружаться данные.

(таблицы для данных из файлов CSV были созданы ранее в практике 4_2, ниже отражено создание таблицы для внесения данных из Postgres)

```
DROP TABLE IF EXISTS employees;
CREATE TABLE employees(
   id integer NOT NULL PRIMARY  KEY,
   first_name VARCHAR(50) NOT NULL,
   last_name VARCHAR(50) NOT NULL,
   email VARCHAR(100),
   phone_number VARCHAR(20),
   hire_date date NOT NULL,
   job_title VARCHAR(50),
   salary NUMERIC(10,2),
   department_id INT
);
```



Создание модуля трансформации в PDI для загрузки данных из Postgres в MySQL:



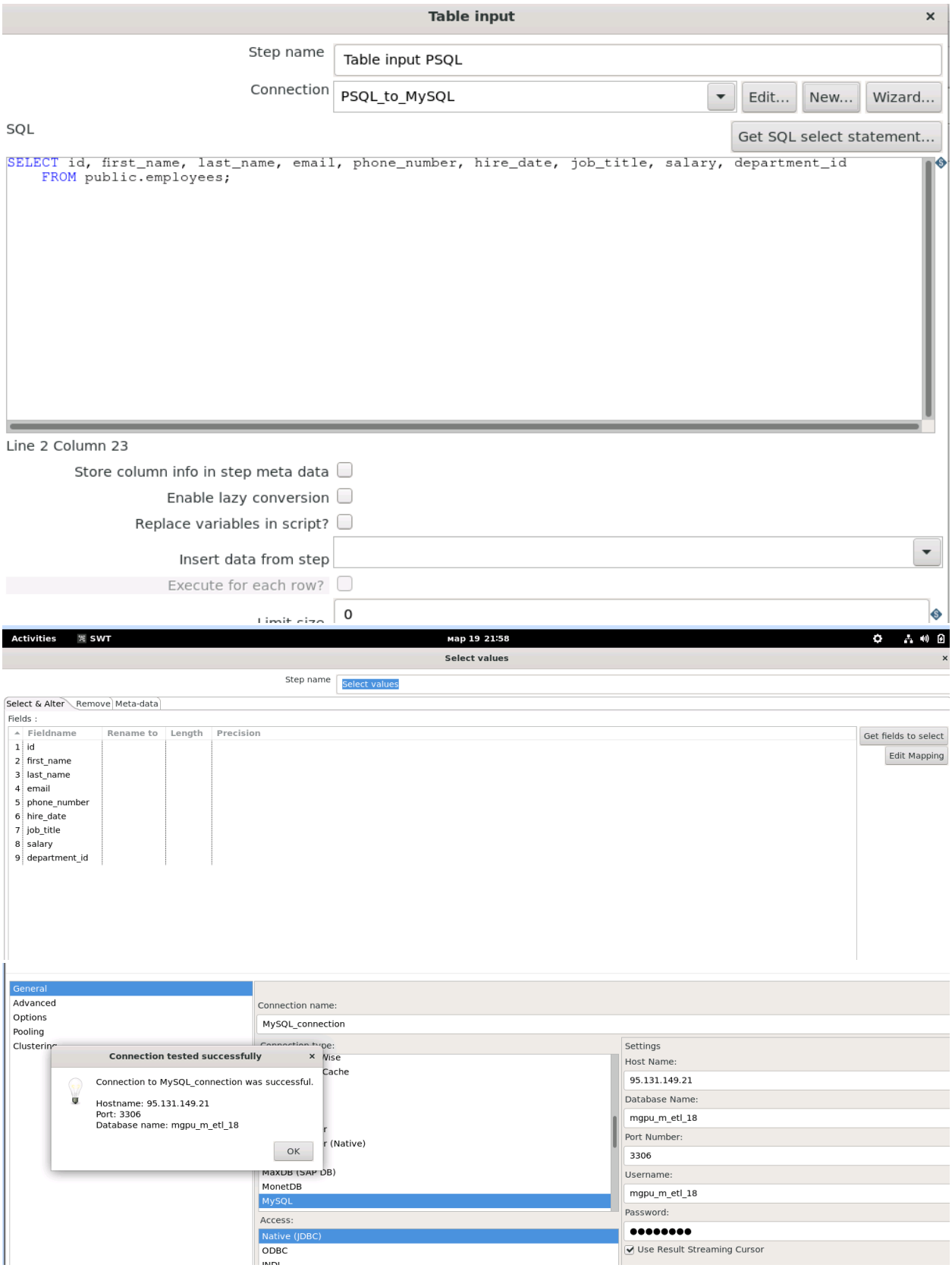
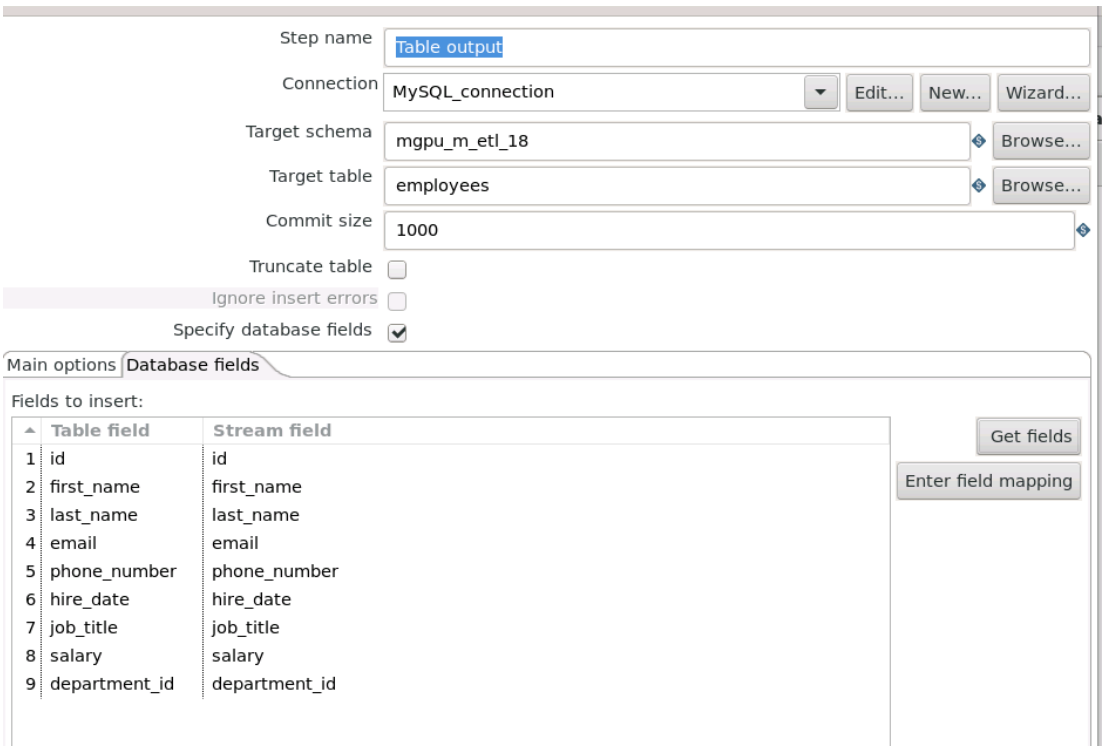
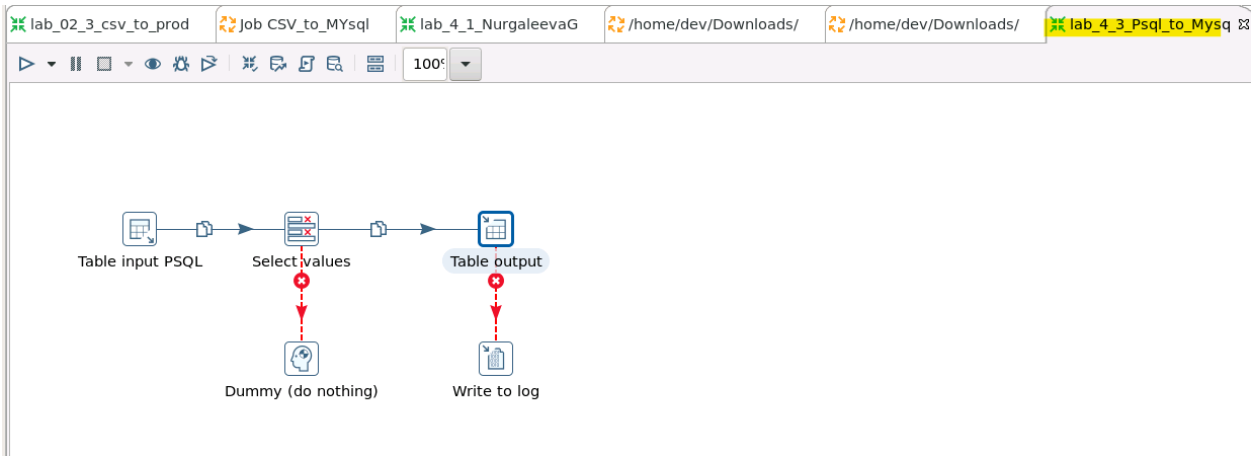
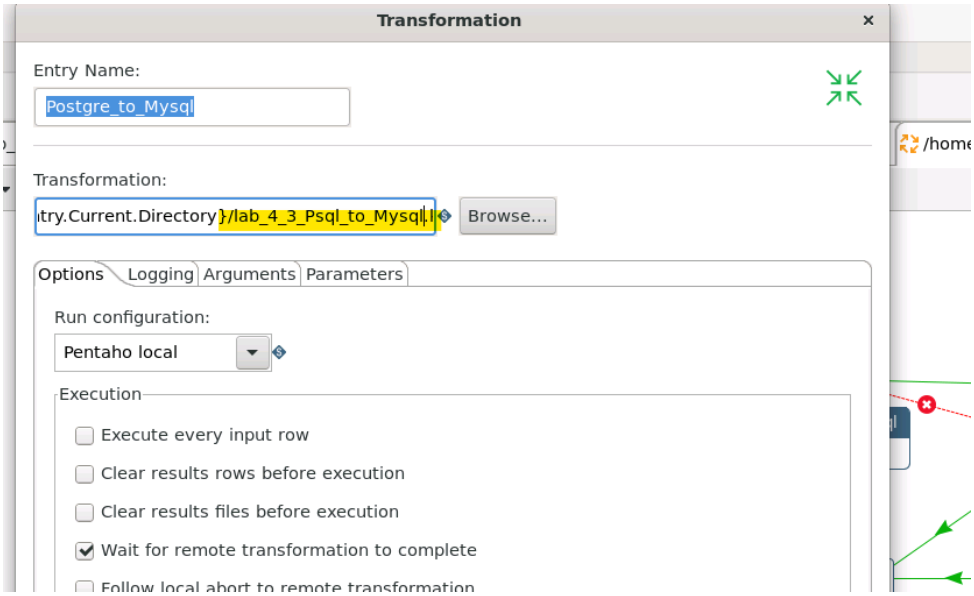


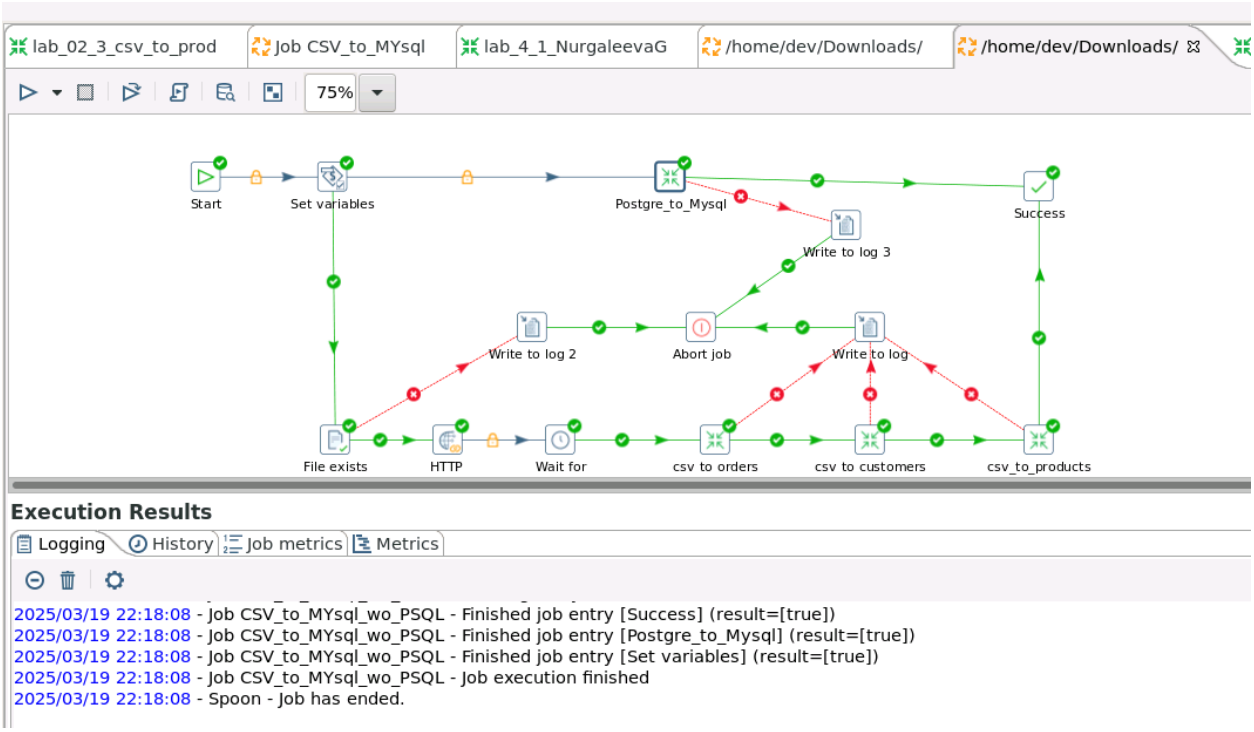
Table input
Step name
Table input PSQL
Connection
PSQL_to_MySQL
Edit...
New...
Wizard...
SQL
Get SQL select statement...
SELECT id, first_name, last_name, email, phone_number, hire_date, job_title, salary, department_id
FROM public.employees;
Line 2 Column 23
Store column info in step meta data
Enable lazy conversion
Replace variables in script?
Insert data from step
Execute for each row?
0
Activities
SWT
мар 19 21:58
Select values
Step name
Select values
Select & Alter
Remove
Meta-data
Fields :
Fieldname
Rename to
Length
Precision
1 id
2 first_name
3 last_name
4 email
5 phone_number
6 hire_date
7 job_title
8 salary
9 department_id
Get fields to select
Edit Mapping
General
Advanced
Options
Pooling
Clustering
Connection name:
MySQL_connection
Connection tested successfully
Connection to MySQL_connection was successful.
Hostname: 95.131.149.21
Port: 3306
Database name: mgpu_m_etl_18
OK
MonetDB
MySQL
Access:
Native (JDBC)
ODBC
Settings
Host Name:
95.131.149.21
Database Name:
mgpu_m_etl_18
Port Number:
3306
Username:
mgpu_m_etl_18
Password:
Use Result Streaming Cursor

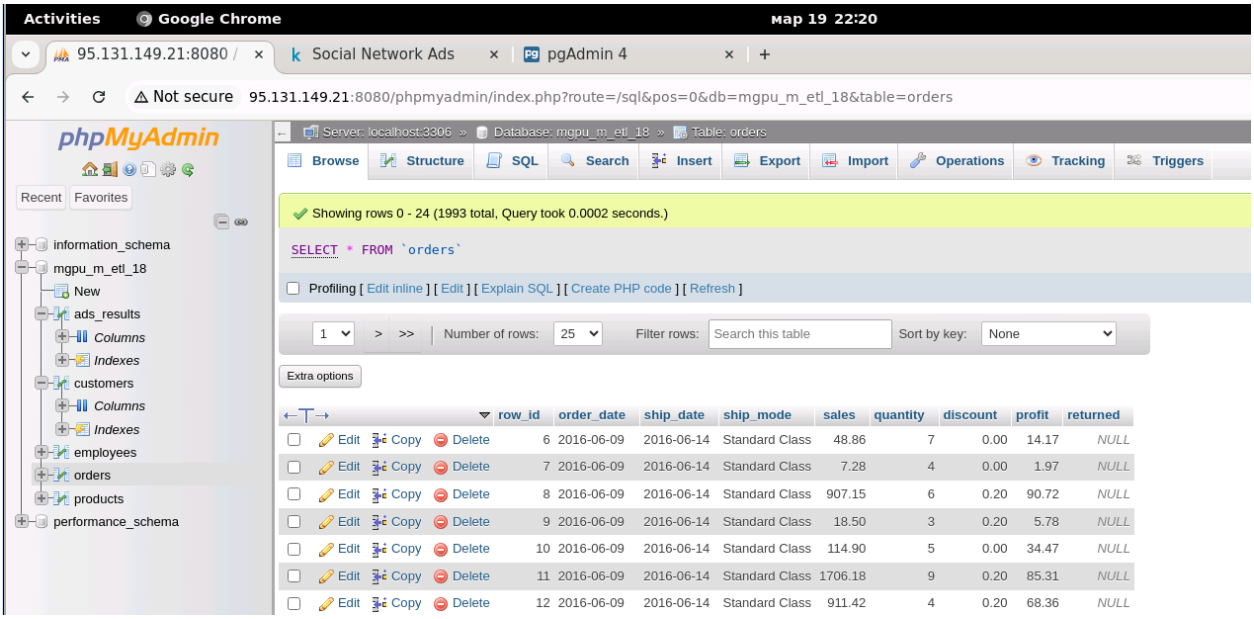

Step name
Table output
Connection
MySQL_connection
Edit...
New...
Wizard...
Target schema
mgpu_m_etl_18
Browse...
Target table
employees
Browse...
Commit size
1000
Truncate table
Ignore insert errors
Specify database fields
Main options
Database fields
Fields to insert:
Table field
Stream field
1 id id
2 first_name first_name
3 last_name last_name
4 email email
5 phone_number phone_number
6 hire_date hire_date
7 job_title job_title
8 salary salary
9 department_id department_id
Get fields
Enter field mapping

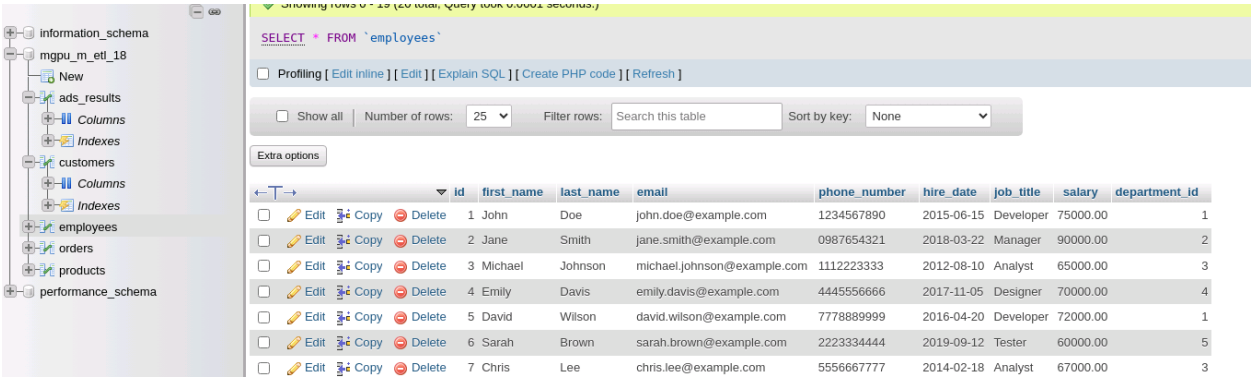

lab_02_3_csv_to_prod
Job CSV_to_MYsql
lab_4_1_NurgaleevaG
/home/dev/Downloads/
/home/dev/Downloads/
lab_4_3_Psql_to_Mysq
Table input PSQL
Select values
Table output
Dummy (do nothing)
Write to log

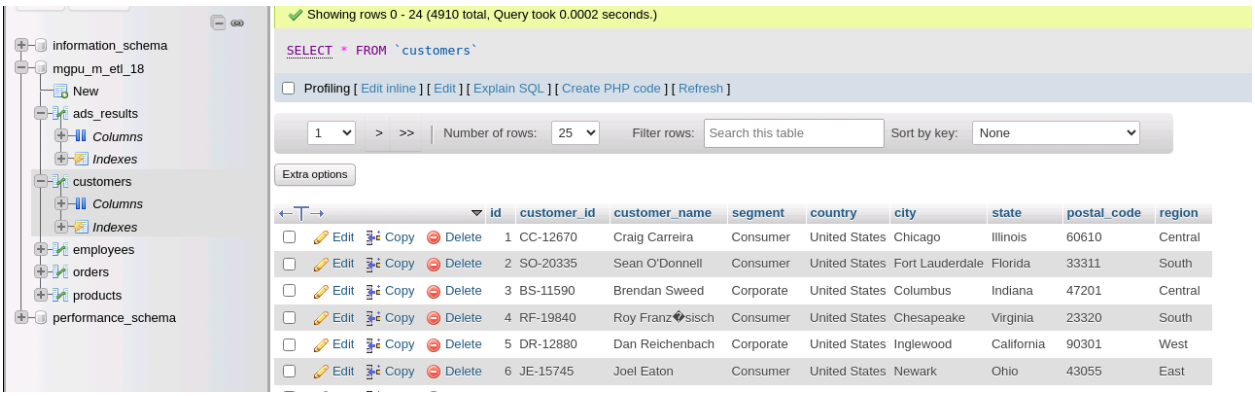

Transformation
Entry Name:
Postgre_to_Mysql
Transformation:
itry.Current.Directory}/lab_4_3_Psql_to_Mysql.ktr
Browse...
Options
Logging
Arguments
Parameters
Run configuration:
Pentaho local
Execution
Execute every input row
Clear results rows before execution
Clear results files before execution
Wait for remote transformation to complete
Follow local abort to remote transformation

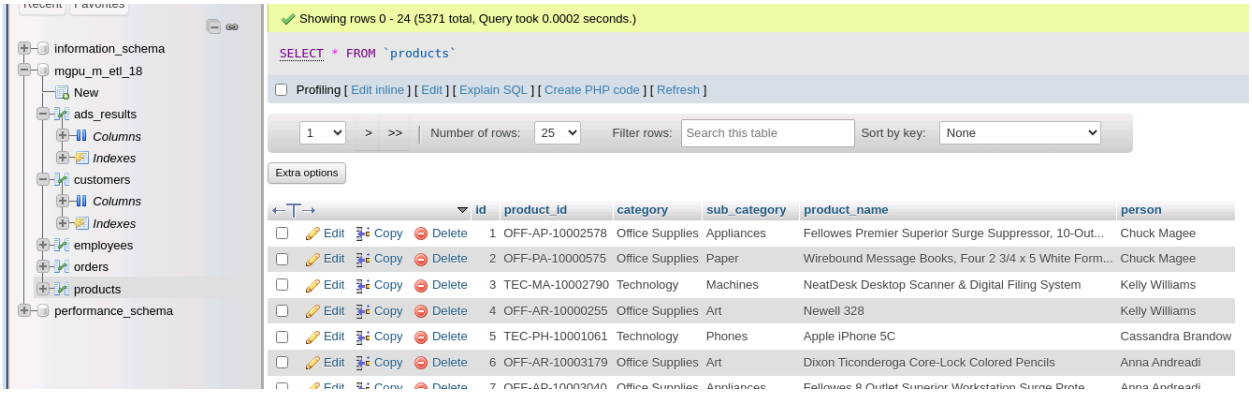
Проверяем, что все нужные таблицы в БД, в которую будет производиться загрузка, существуют и они пустые. При необходимости очищаем.








information_schema
mgpu_m_etl_18
New
ads_results
Columns
Indexes
customers
Columns
Indexes
employees
orders
products
performance_schema
Showing rows 0 - 24 (4910 total, Query took 0.0002 seconds.)
SELECT * FROM `customers`
Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]
Number of rows: 25
Filter rows: Search this table
Sort by key: None
Extra options
id customer_id customer_name segment country city state postal_code region
Edit Copy Delete 1 CC-12670 Craig Carreira Consumer United States Chicago Illinois 60610 Central
Edit Copy Delete 2 SO-20335 Sean O'Donnell Consumer United States Fort Lauderdale Florida 33311 South
Edit Copy Delete 3 BS-11590 Brendan Sweed Corporate United States Columbus Indiana 47201 Central
Edit Copy Delete 4 RF-19840 Roy Franz?sisch Consumer United States Chesapeake Virginia 23320 South
Edit Copy Delete 5 DR-12880 Dan Reichenbach Corporate United States Inglewood California 90301 West
Edit Copy Delete 6 JE-15745 Joel Eaton Consumer United States Newark Ohio 43055 East


information_schema
mgpu_m_etl_18
New
ads_results
Columns
Indexes
customers
Columns
Indexes
employees
orders
products
performance_schema
Showing rows 0 - 24 (5371 total, Query took 0.0002 seconds.)
SELECT * FROM `products`
Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]
Number of rows: 25
Filter rows: Search this table
Sort by key: None
Extra options
id product_id category sub_category product_name person
Edit Copy Delete 1 OFF-AP-10002578 Office Supplies Appliances Fellowes Premier Superior Surge Suppressor, 10-Out... Chuck Magee
Edit Copy Delete 2 OFF-PA-10000575 Office Supplies Paper Wirebound Message Books, Four 2 3/4 x 5 White Form... Chuck Magee
Edit Copy Delete 3 TEC-MA-10002790 Technology Machines NeatDesk Desktop Scanner & Digital Filing System Kelly Williams
Edit Copy Delete 4 OFF-AR-10000255 Office Supplies Art Newell 328 Kelly Williams
Edit Copy Delete 5 TEC-PH-10001061 Technology Phones Apple iPhone 5C Cassandra Brandow
Edit Copy Delete 6 OFF-AR-10003179 Office Supplies Art Dixon Ticonderoga Core-Lock Colored Pencils Anna Andreadi