

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),
```

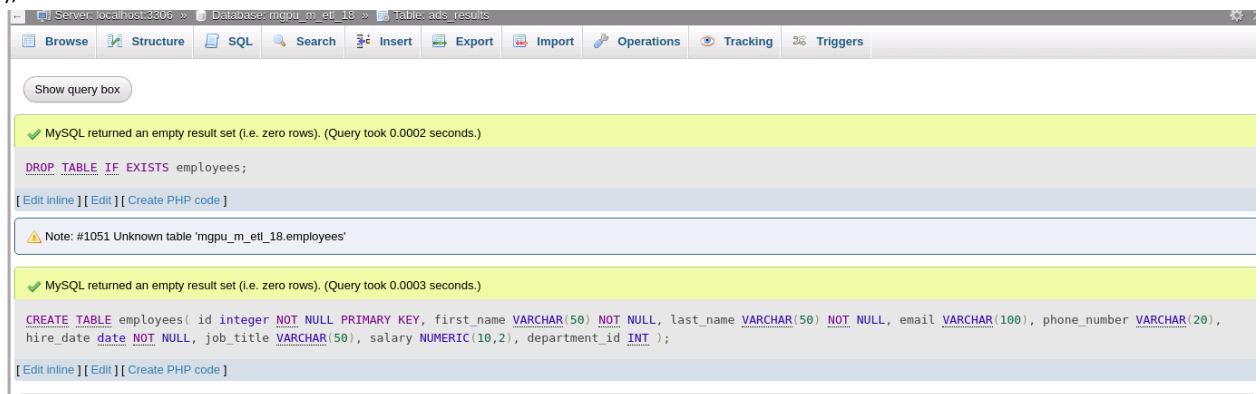
```
    hire_date date NOT NULL,
```

```
    job_title VARCHAR(50),
```

```
    salary NUMERIC(10,2),
```

```
    department_id INT
```

```
);
```



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

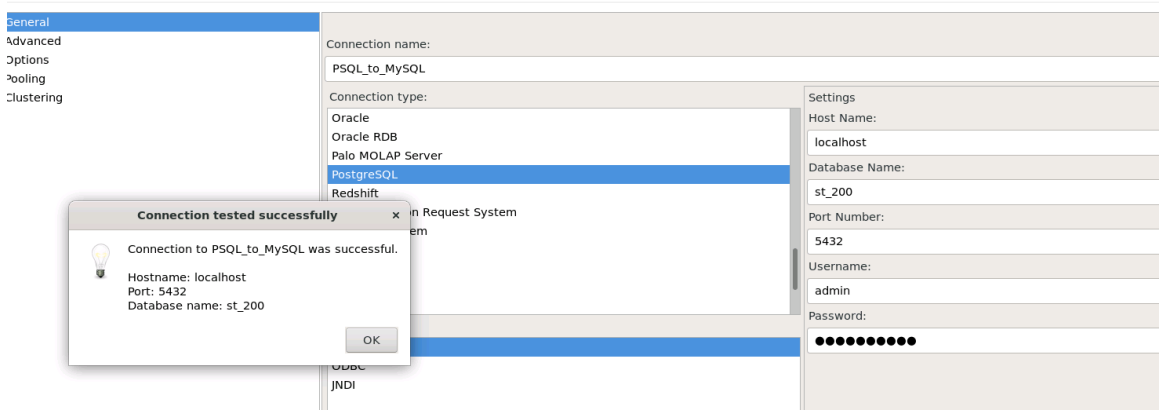


Table input

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?

Limit size

Activities SWTmap 19 21:58

Select values

Step name

Select values

Select & AlterRemoveMeta-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

GeneralAdvancedOptionsPoolingClustering

Connection name:

MySQL_connection

Connection type:

Cache

(Native)

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

Connection tested successfully

Connection to MySQL_connection was successful.

Hostname: 95.131.149.21

Port: 3306

Database name: mgpu_m_etl_18

OK

Step name:

Connection: Edit... New... Wizard...

Target schema: Browse...

Target table: Browse...

Commit size:

Truncate table: ☐

Ignore insert errors: ☐

Specify database fields: ☒

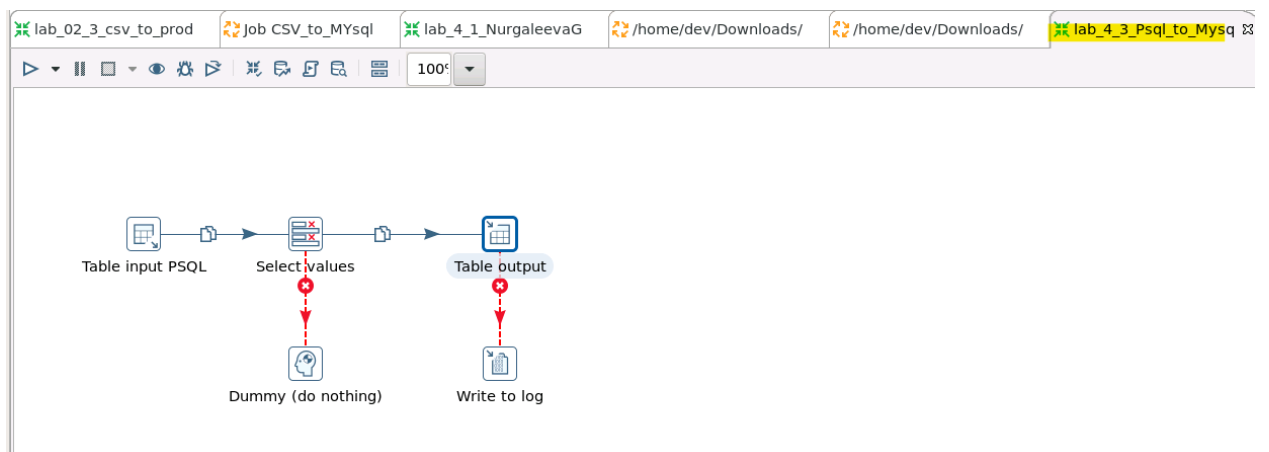
Main options | Database fields

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



Transformation

Entry Name:

Transformation: Browse...

Options | Logging | Arguments | Parameters

Run configuration:

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.

lab_02_3_csv_to_prod Job CSV_to_Mysql lab_4_1_NurgaleevaG /home/dev/Downloads/ /home/dev/Downloads/ 75%

Execution Results

Logging History Job metrics Metrics

2025/03/19 22:18:08 - Job CSV_to_Mysql_wo_PSQL - Finished job entry [Success] (result=[true])
 2025/03/19 22:18:08 - Job CSV_to_Mysql_wo_PSQL - Finished job entry [Postgre_to_Mysql] (result=[true])
 2025/03/19 22:18:08 - Job CSV_to_Mysql_wo_PSQL - Finished job entry [Set variables] (result=[true])
 2025/03/19 22:18:08 - Job CSV_to_Mysql_wo_PSQL - Job execution finished
 2025/03/19 22:18:08 - Spoon - Job has ended.

Activities Google Chrome map 19 22:20

95.131.149.21:8080 / x Social Network Ads x pgAdmin 4 x +

Not secure 95.131.149.21:8080/phpmyadmin/index.php?route=/sql&pos=0&db=mgpu_m_etl_18&table=orders

Server: localhost:3306 Database: mgpu_m_etl_18 Table: orders

Browse Structure SQL Search Insert Export Import Operations Tracking Triggers

Showing rows 0 - 24 (1993 total, Query took 0.0002 seconds.)

SELECT * FROM `orders`

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

	row_id	order_date	ship_date	ship_mode	sales	quantity	discount	profit	returned
<input type="checkbox"/>	6	2016-06-09	2016-06-14	Standard Class	48.86	7	0.00	14.17	NULL
<input type="checkbox"/>	7	2016-06-09	2016-06-14	Standard Class	7.28	4	0.00	1.97	NULL
<input type="checkbox"/>	8	2016-06-09	2016-06-14	Standard Class	907.15	6	0.20	90.72	NULL
<input type="checkbox"/>	9	2016-06-09	2016-06-14	Standard Class	18.50	3	0.20	5.78	NULL
<input type="checkbox"/>	10	2016-06-09	2016-06-14	Standard Class	114.90	5	0.00	34.47	NULL
<input type="checkbox"/>	11	2016-06-09	2016-06-14	Standard Class	1706.18	9	0.20	85.31	NULL
<input type="checkbox"/>	12	2016-06-09	2016-06-14	Standard Class	911.42	4	0.20	68.36	NULL

Extra options

Showing rows 0 - 24 (1993 total, Query took 0.0002 seconds.)

SELECT * FROM `employees`

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

Show all Number of rows: 25 Filter rows: Search this table Sort by key: None

Extra options

	id	first_name	last_name	email	phone_number	hire_date	job_title	salary	department_id
<input type="checkbox"/>	1	John	Doe	john.doe@example.com	1234567890	2015-06-15	Developer	75000.00	1
<input type="checkbox"/>	2	Jane	Smith	jane.smith@example.com	0987654321	2018-03-22	Manager	90000.00	2
<input type="checkbox"/>	3	Michael	Johnson	michael.johnson@example.com	1112223333	2012-08-10	Analyst	65000.00	3
<input type="checkbox"/>	4	Emily	Davis	emily.davis@example.com	4445556666	2017-11-05	Designer	70000.00	4
<input type="checkbox"/>	5	David	Wilson	david.wilson@example.com	7778889999	2016-04-20	Developer	72000.00	1
<input type="checkbox"/>	6	Sarah	Brown	sarah.brown@example.com	2223334444	2019-09-12	Tester	60000.00	5
<input type="checkbox"/>	7	Chris	Lee	chris.lee@example.com	5556667777	2014-02-18	Analyst	67000.00	3

Extra options

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]

1 > >> 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		
<input type="checkbox"/>	Edit	Copy	Delete	1	CC-12670	Craig Carreira	Consumer	United States	Chicago	Illinois	60610	Central
<input type="checkbox"/>	Edit	Copy	Delete	2	SO-20335	Sean O'Donnell	Consumer	United States	Fort Lauderdale	Florida	33311	South
<input type="checkbox"/>	Edit	Copy	Delete	3	BS-11590	Brendan Sweed	Corporate	United States	Columbus	Indiana	47201	Central
<input type="checkbox"/>	Edit	Copy	Delete	4	RF-19840	Roy Franz	Consumer	United States	Chesapeake	Virginia	23320	South
<input type="checkbox"/>	Edit	Copy	Delete	5	DR-12880	Dan Reichenbach	Corporate	United States	Inglewood	California	90301	West
<input type="checkbox"/>	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]

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

Extra options

		id	product_id	category	sub_category	product_name	person	
<input type="checkbox"/>	Edit	Copy	Delete	1	OFF-AP-10002578	Office Supplies Appliances	Fellowes Premier Superior Surge Suppressor, 10-Out...	Chuck Magee
<input type="checkbox"/>	Edit	Copy	Delete	2	OFF-PA-10000575	Office Supplies Paper	Wirebound Message Books, Four 2 3/4 x 5 White Form...	Chuck Magee
<input type="checkbox"/>	Edit	Copy	Delete	3	TEC-MA-10002790	Technology Machines	NeatDesk Desktop Scanner & Digital Filing System	Kelly Williams
<input type="checkbox"/>	Edit	Copy	Delete	4	OFF-AR-10000255	Office Supplies Art	Newell 328	Kelly Williams
<input type="checkbox"/>	Edit	Copy	Delete	5	TEC-PH-10001061	Technology Phones	Apple iPhone 5C	Cassandra Brandow
<input type="checkbox"/>	Edit	Copy	Delete	6	OFF-AR-10003179	Office Supplies Art	Dixon Ticonderoga Core-Lock Colored Pencils	Anna Andreadi
<input type="checkbox"/>	Edit	Copy	Delete	7	OFF-AP-10003040	Office Supplies Appliances	Fellowes 8 Outlet Surge Protector Surge Protector	Anna Andreadi