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

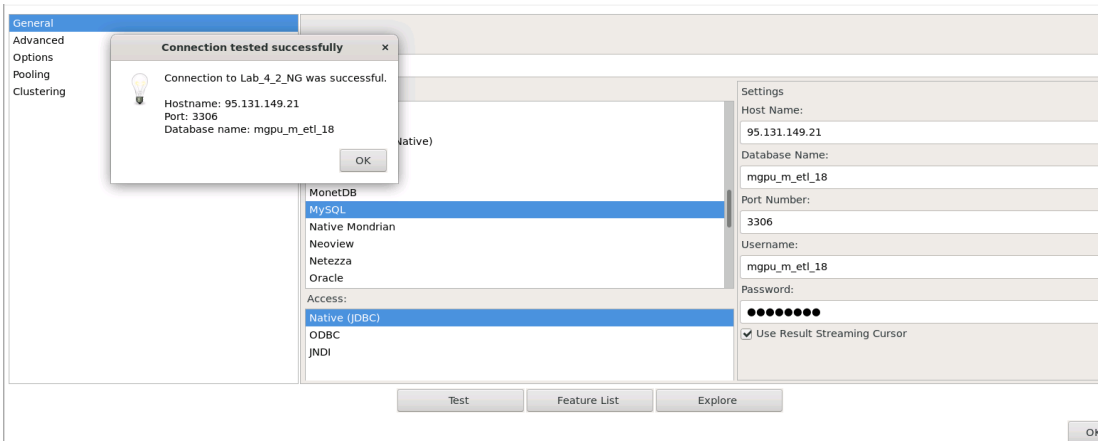
Нургалеева Гузель Рустэмовна БД-241м

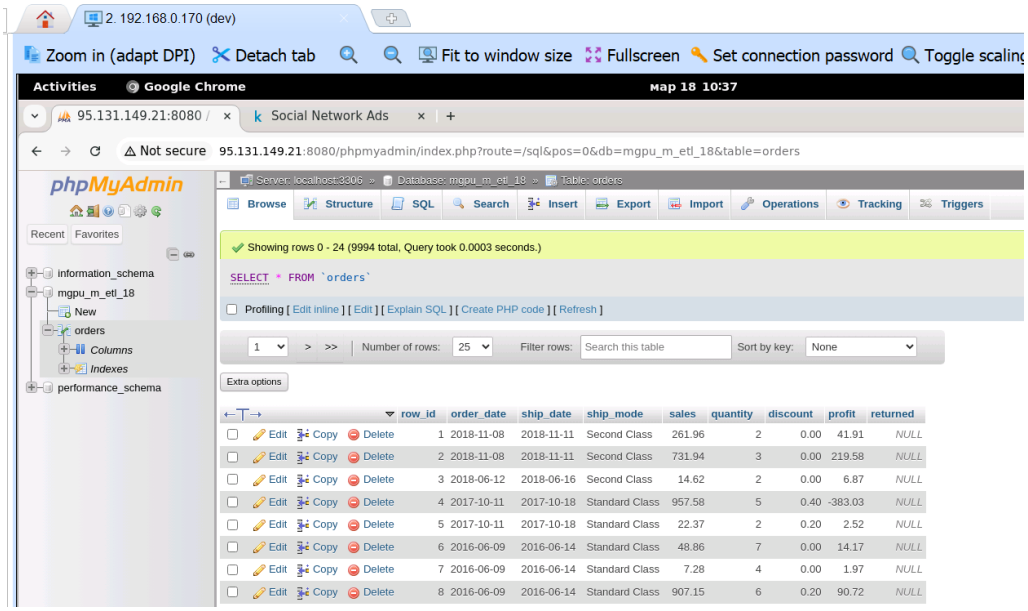
Data Engineering. Практическая работа 4-2.
Вариант задания: 18

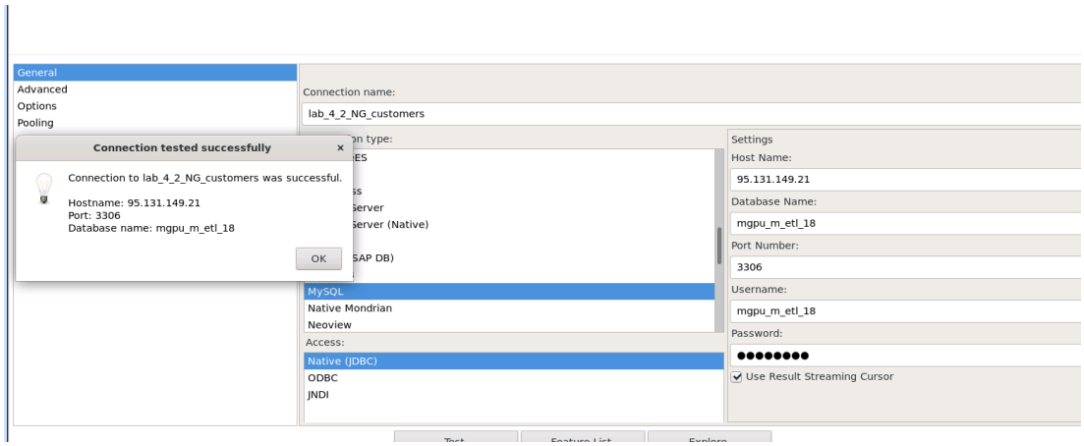
Направление подготовки/специальность
38.04.05 - Бизнес-информатика
Бизнес-аналитика и большие данные
(очная форма обучения)

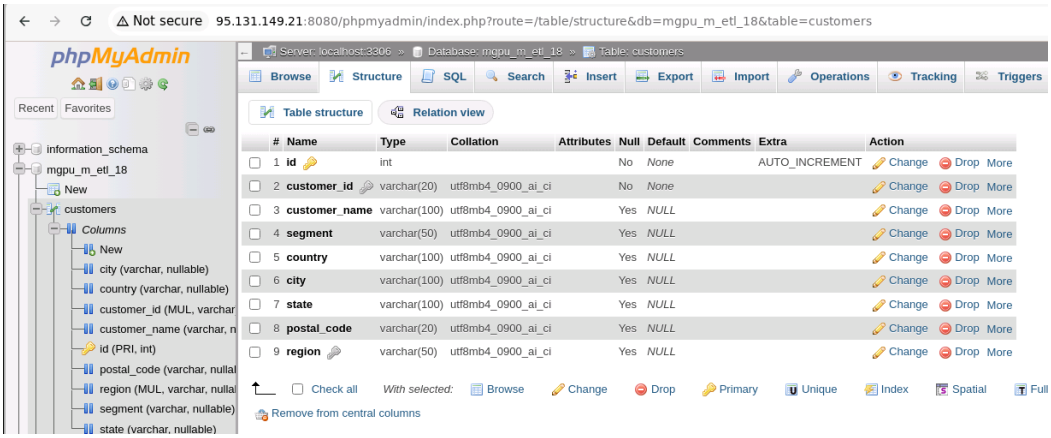
Москва
2024

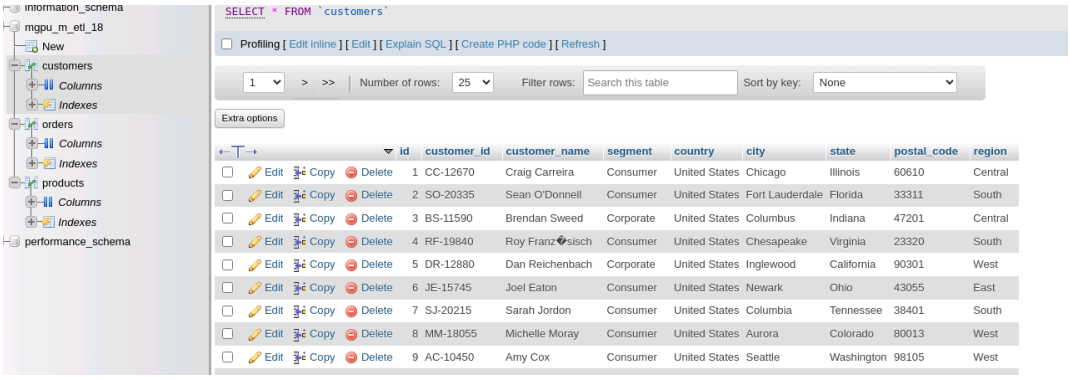
Цель работы: получить практические навыки создания ETL-процесса для загрузки данных из CSV-файла в базу данных MySQL с использованием Pentaho Data Integration.

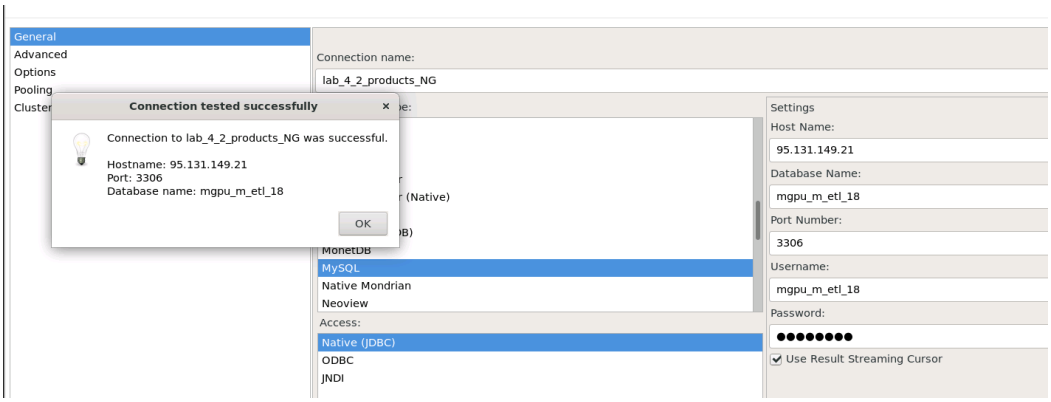


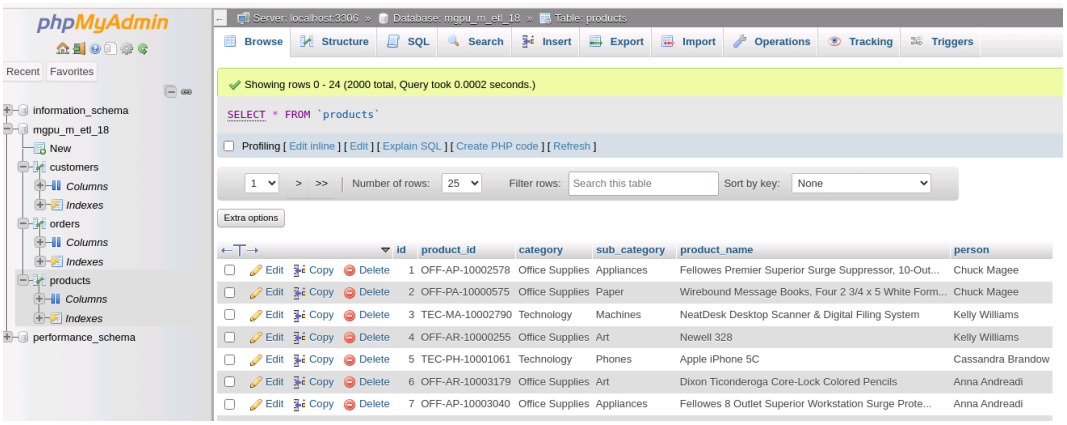




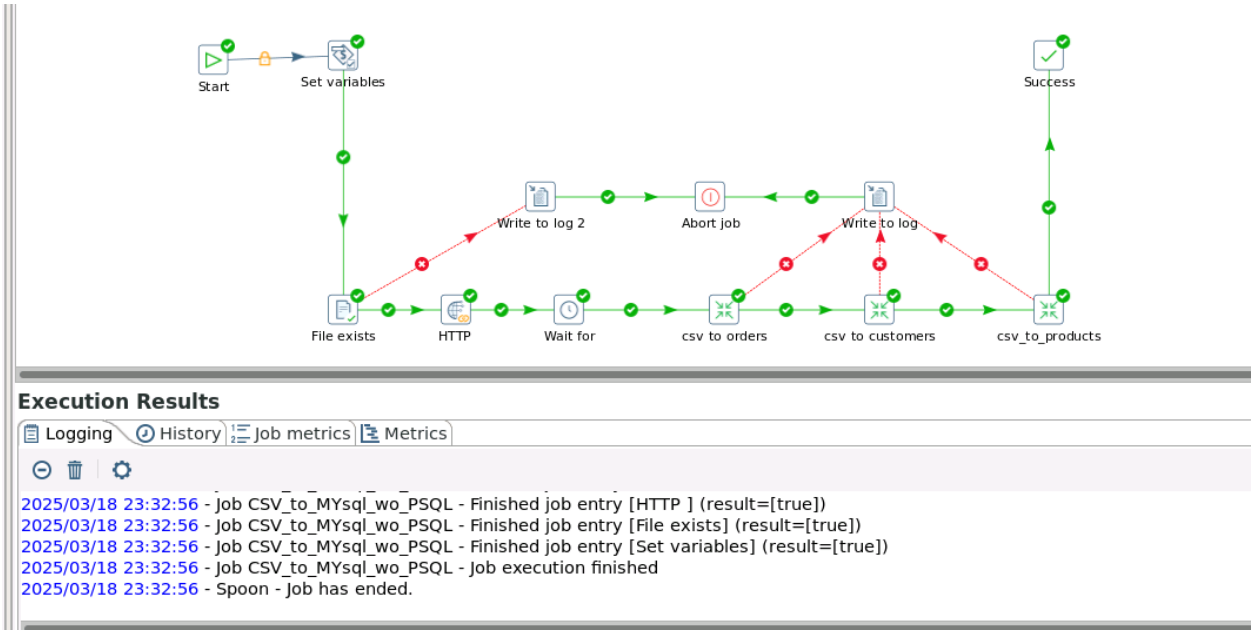

Not secure 95.131.149.21:8080/phpmyadmin/index.php?route=/table/structure&db=mgpu_m_etl_18&table=customers
phpMyAdmin
Server: localhost:3306 » Database: mgpu_m_etl_18 » Table: customers
Browse Structure SQL Search Insert Export Import Operations Tracking Triggers
Table structure Relation view
Name Type Collation Attributes Null Default Comments Extra Action
1 id int No None AUTO_INCREMENT Change Drop More
2 customer_id varchar(20) utf8mb4_0900_ai_ci No None Change Drop More
3 customer_name varchar(100) utf8mb4_0900_ai_ci Yes NULL Change Drop More
4 segment varchar(50) utf8mb4_0900_ai_ci Yes NULL Change Drop More
5 country varchar(100) utf8mb4_0900_ai_ci Yes NULL Change Drop More
6 city varchar(100) utf8mb4_0900_ai_ci Yes NULL Change Drop More
7 state varchar(100) utf8mb4_0900_ai_ci Yes NULL Change Drop More
8 postal_code varchar(20) utf8mb4_0900_ai_ci Yes NULL Change Drop More
9 region varchar(50) utf8mb4_0900_ai_ci Yes NULL Change Drop More
Check all With selected: Browse Change Drop Primary Unique Index Spatial Full
Remove from central columns

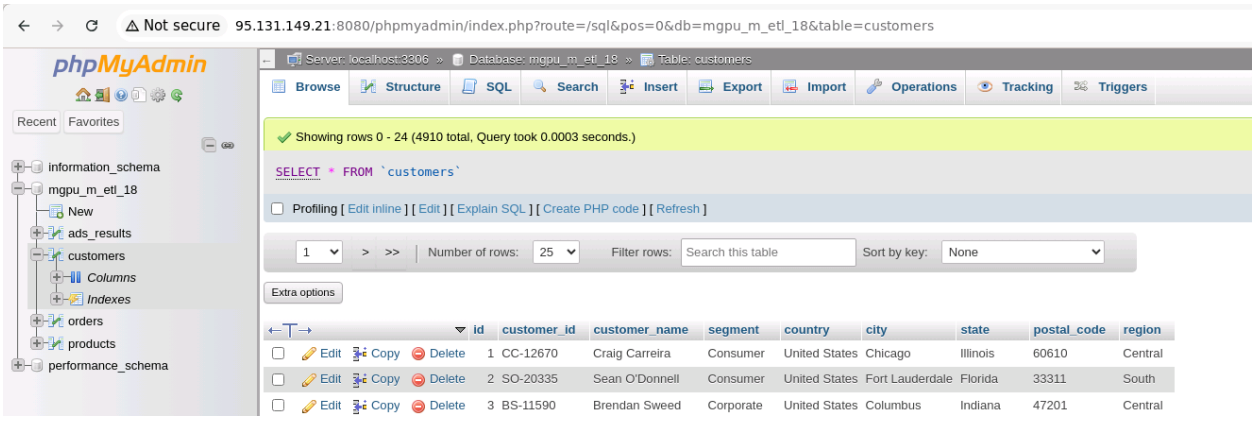

SELECT * FROM `customers`
Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]
Number of rows: 25 Filter rows: Search this table Sort by key: None
Extra options
id customer_id customer_name segment country city state postal_code region
1 CC-12670 Craig Carreira Consumer United States Chicago Illinois 60610 Central
2 SO-20335 Sean O'Donnell Consumer United States Fort Lauderdale Florida 33311 South
3 BS-11590 Brendan Sweed Corporate United States Columbus Indiana 47201 Central
4 RF-19840 Roy Franz?sisch Consumer United States Chesapeake Virginia 23320 South
5 DR-12880 Dan Reichenbach Corporate United States Inglewood California 90301 West
6 JE-15745 Joel Eaton Consumer United States Newark Ohio 43055 East
7 SJ-20215 Sarah Jordon Consumer United States Columbia Tennessee 38401 South
8 MM-18055 Michelle Moray Consumer United States Aurora Colorado 80013 West
9 AC-10450 Amy Cox Consumer United States Seattle Washington 98105 West

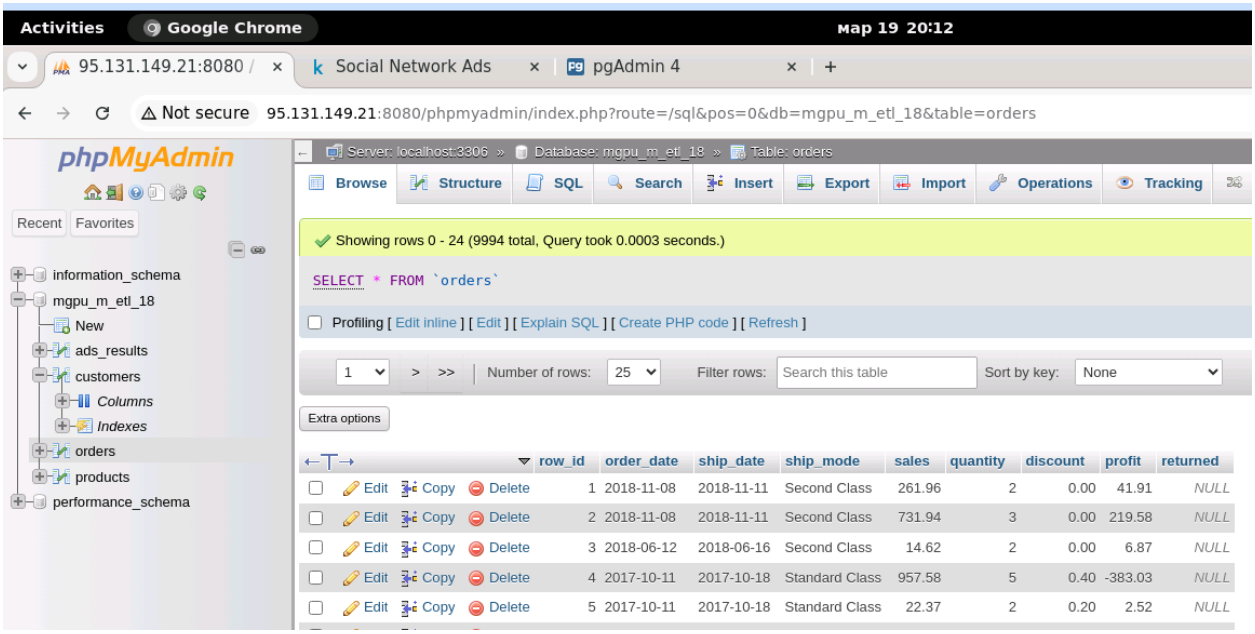

General
Advanced
Options
Pooling
Connection tested successfully
Connection to lab_4_2_products_NG was successful.
Hostname: 95.131.149.21
Port: 3306
Database name: mgpu_m_etl_18
OK
Connection name:
lab_4_2_products_NG
MonetDB
MySQL
Native Mondrian
Neoview
Access:
Native (JDBC)
ODBC
JNDI
Settings
Host Name:
95.131.149.21
Database Name:
mgpu_m_etl_18
Port Number:
3306
Username:
mgpu_m_etl_18
Password:
Use Result Streaming Cursor

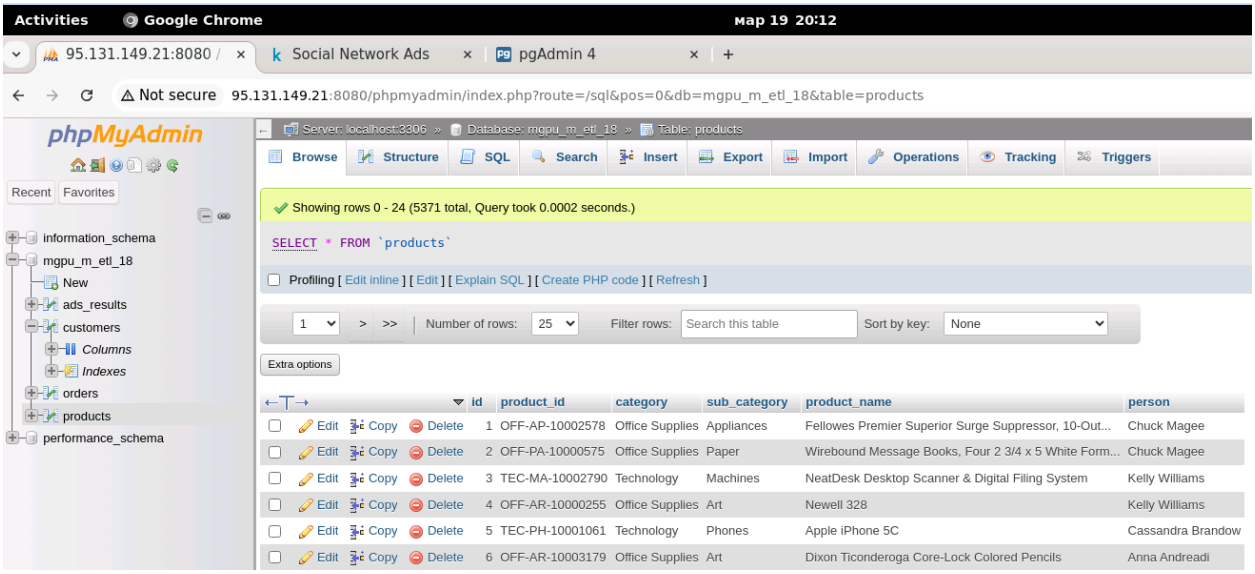

phpMyAdmin
Server: localhost:3306 » Database: mgpu_m_etl_18 » Table: products
Browse Structure SQL Search Insert Export Import Operations Tracking Triggers
Showing rows 0 - 24 (2000 total, Query took 0.0002 seconds.)
SELECT * FROM `products`
Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [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
1 OFF-AP-10002578 Office Supplies Appliances Fellowes Premier Superior Surge Suppressor, 10-Out... Chuck Magee
2 OFF-PA-10000575 Office Supplies Paper Wirebound Message Books, Four 2 3/4 x 5 White Form... Chuck Magee
3 TEC-MA-10002790 Technology Machines NeatDesk Desktop Scanner & Digital Filing System Kelly Williams
4 OFF-AR-10000255 Office Supplies Art Newell 328 Kelly Williams
5 TEC-PH-10001061 Technology Phones Apple iPhone 5C Cassandra Brandow
6 OFF-AR-10003179 Office Supplies Art Dixon Ticonderoga Core-Lock Colored Pencils Anna Andreadi
7 OFF-AP-10003040 Office Supplies Appliances Fellowes 8 Outlet Superior Workstation Surge Prote... Anna Andreadi

И вариант с использованием job

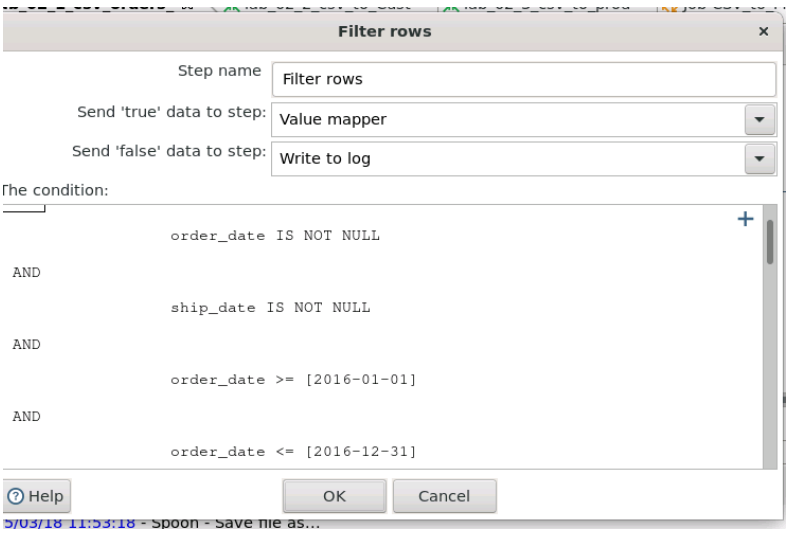


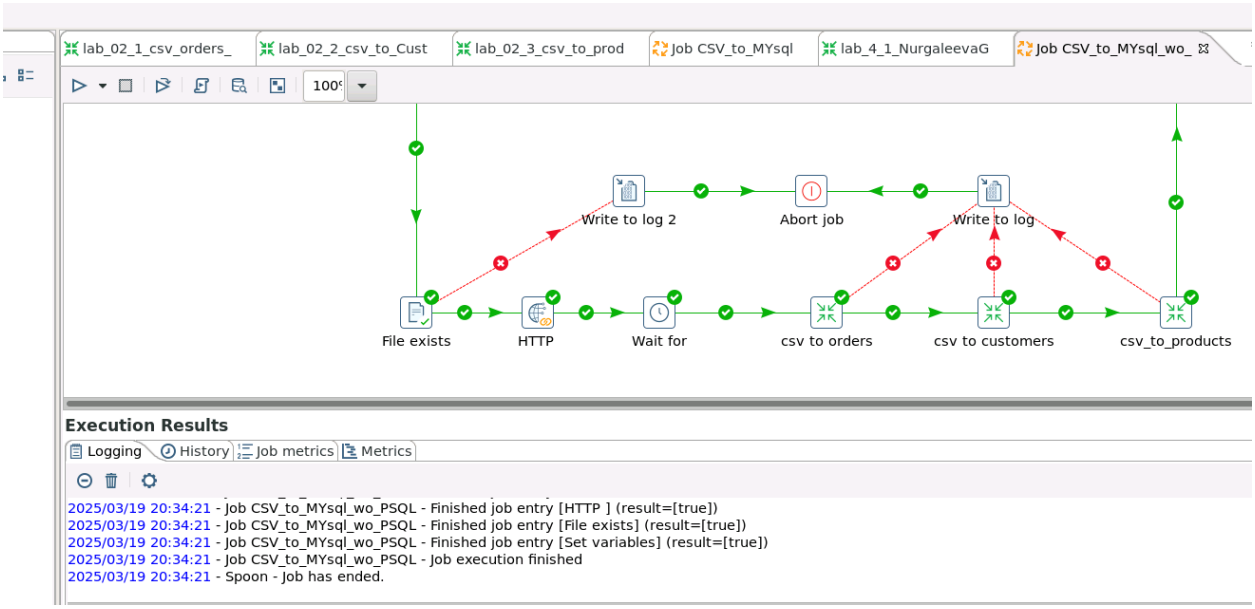


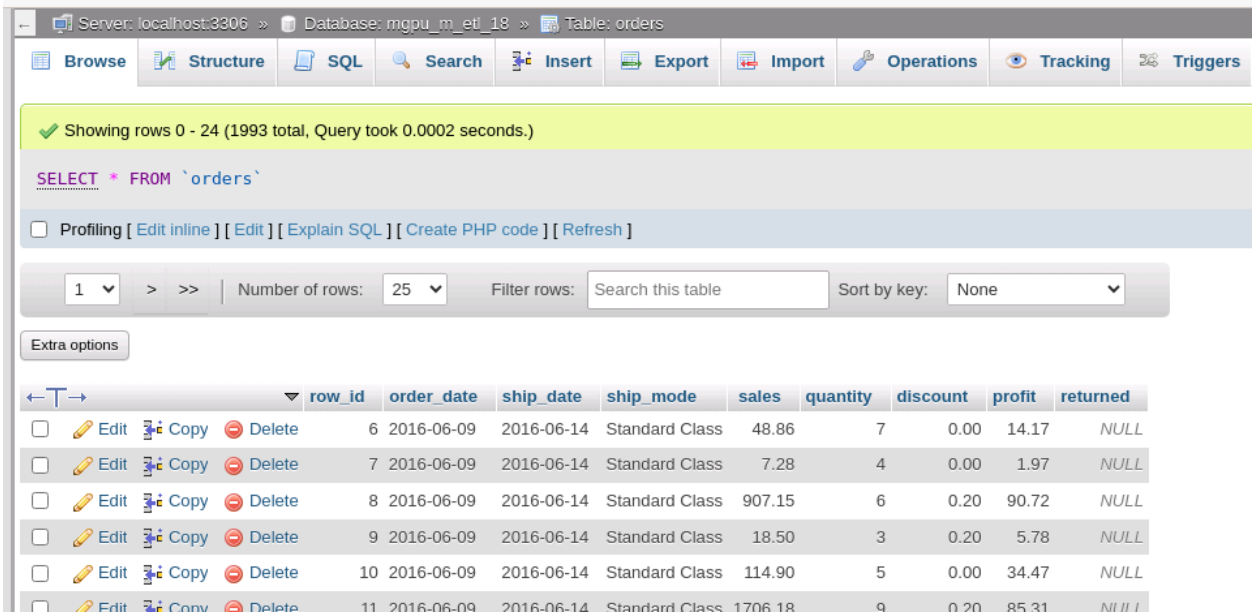




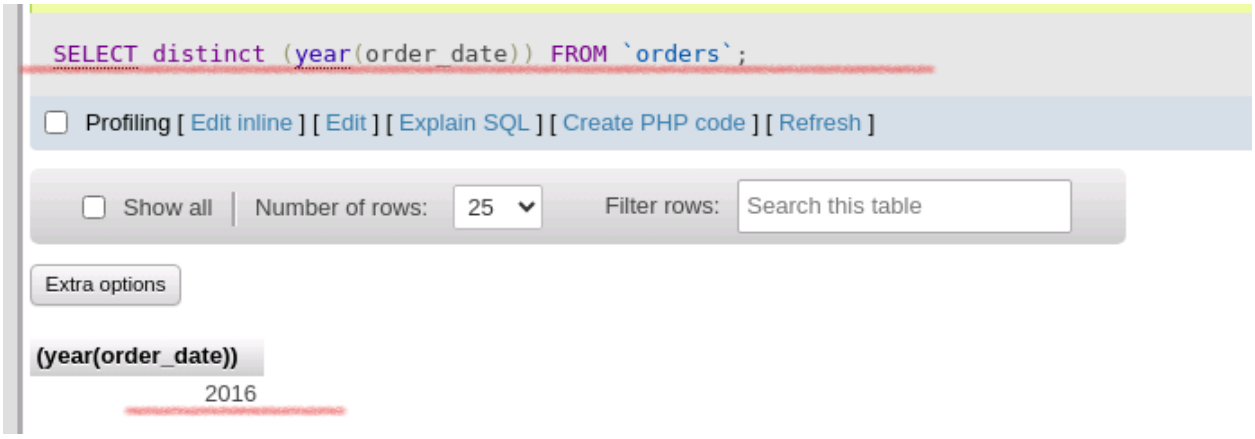
Индивидуальное задание 18: загрузить данные с фильтром по дате – только заказы 2016 года




lab_02_1_csv_orders_
lab_02_2_csv_to_Cust
lab_02_3_csv_to_prod
Job CSV_to_MYsql
lab_4_1_NurgaleevaG
Job CSV_to_MYsql_wo_
Write to log 2
Abort job
Write to log
File exists
HTTP
Wait for
csv to orders
csv to customers
csv_to_products
Execution Results
Logging
History
Job metrics
Metrics
2025/03/19 20:34:21 - Job CSV_to_MYsql_wo_PSQL - Finished job entry [HTTP] (result=[true])
2025/03/19 20:34:21 - Job CSV_to_MYsql_wo_PSQL - Finished job entry [File exists] (result=[true])
2025/03/19 20:34:21 - Job CSV_to_MYsql_wo_PSQL - Finished job entry [Set variables] (result=[true])
2025/03/19 20:34:21 - Job CSV_to_MYsql_wo_PSQL - Job execution finished
2025/03/19 20:34:21 - Spoon - Job has ended.


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]
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
Edit Copy Delete 6 2016-06-09 2016-06-14 Standard Class 48.86 7 0.00 14.17 NULL
Edit Copy Delete 7 2016-06-09 2016-06-14 Standard Class 7.28 4 0.00 1.97 NULL
Edit Copy Delete 8 2016-06-09 2016-06-14 Standard Class 907.15 6 0.20 90.72 NULL
Edit Copy Delete 9 2016-06-09 2016-06-14 Standard Class 18.50 3 0.20 5.78 NULL
Edit Copy Delete 10 2016-06-09 2016-06-14 Standard Class 114.90 5 0.00 34.47 NULL

Проверка:


SELECT distinct (year(order_date)) FROM `orders`;
Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]
Show all
Number of rows: 25
Filter rows: Search this table
Extra options
(year(order_date))
2016

SQL скрипты

```
-- Таблица заказов (основная информация о продажах)
CREATE TABLE orders (
row_id INT PRIMARY KEY,
order_date DATE,
ship_date DATE,
ship_mode VARCHAR(50),
sales DECIMAL(10,2),
quantity INT,
discount DECIMAL(4,2),
profit DECIMAL(10,2),
returned TINYINT(1) DEFAULT 0 -- 1 = Yes, 0 = No
);

-- Таблица клиентов
DROP TABLE IF EXISTS customers;
CREATE TABLE customers (
id INT AUTO_INCREMENT PRIMARY KEY,
customer_id VARCHAR(20) NOT NULL,
customer_name VARCHAR(100),
segment VARCHAR(50),
country VARCHAR(100),
city VARCHAR(100),
state VARCHAR(100),
postal_code VARCHAR(20),
region VARCHAR(50),
INDEX idx_customer_id (customer_id),
INDEX idx_region (region)
);

-- создаем таблицу products
DROP TABLE IF EXISTS products;
CREATE TABLE products (
id INT AUTO_INCREMENT PRIMARY KEY,
product_id VARCHAR(20) NOT NULL,
category VARCHAR(50),
sub_category VARCHAR(50),
product_name VARCHAR(255),
person VARCHAR(100),
INDEX idx_product_id (product_id),
INDEX idx_category (category),
INDEX idx_subcategory (sub_category)
);
```

Контрольные вопросы

1. **Что такое динамические соединения в PDI?**

Динамические соединения в PDI позволяют:
- Использовать параметры подключения из внешних источников.
- Менять настройки соединения во время выполнения.
- Обрабатывать множество источников данных в одном процессе.

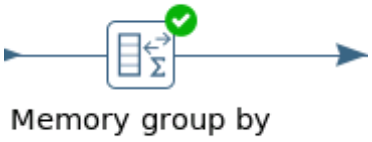
2. **Как организовать обработку ошибок в трансформации?**

Обработку ошибок можно организовать через занесение их в Log. После этого можно просмотреть журнал ошибок.
Также при варианте, в котором задействуется шаг с Log, можно предусмотреть последующим шагом остановку выполнения алгоритма.

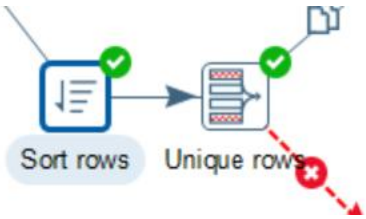
Компоненты обработки ошибок.
- wrt-execution_error - запись информации об ошибках.
- abrt-execution_error - прерывание выполнения при критических ошибках.

3. **Какие методы выявления дублей существуют?**

- можно удалить дубли используя группировку по всем столбцам



- через использование функции unique rows



4. **Как настроить параметризацию подключений?**

Для настройки параметризации подключений в Pentaho можно использовать именованные параметры.
Эта система позволяет параметризовать преобразования и задания.
Чтобы настроить параметры, нужно:

1. Перейти в диалоговое окно настроек преобразования или задания и найти вкладку «Параметры».
2. Установить значения для параметров, которые будут использоваться во время выполнения. Если для параметра не установлено значение, используется значение по умолчанию.
3. В диалоговом окне выполнения преобразований и заданий можно установить значение для каждого определённого именованного параметра.

5. **Какие компоненты PDI используются для объединения данных?**

При помощи компонента **Pentaho Data Integration** также известного как Kettle. Он используется для интеграции данных из разных источников.
Один из востребованных шагов в Pentaho Data Integration — объединение (join) двух потоков данных. Для этого в настройках можно выбрать тип соединения (Inner, Left Outer, Right Outer, Full Outer) и указать ключи.
Обязательное требование для шага join — входные данные должны быть отсортированы по ключевым полям, для этого в Pentaho Data Integration применяется отдельный шаг **Sorter**