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

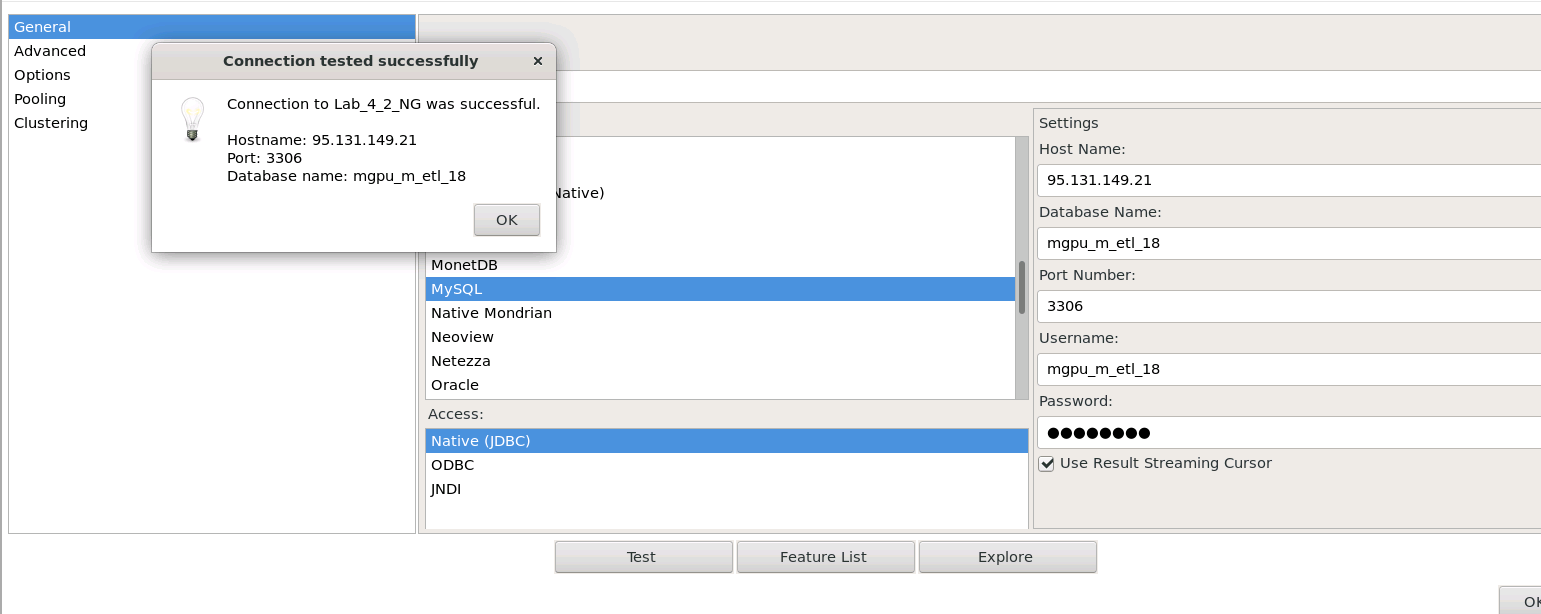
(очная форма обучения)

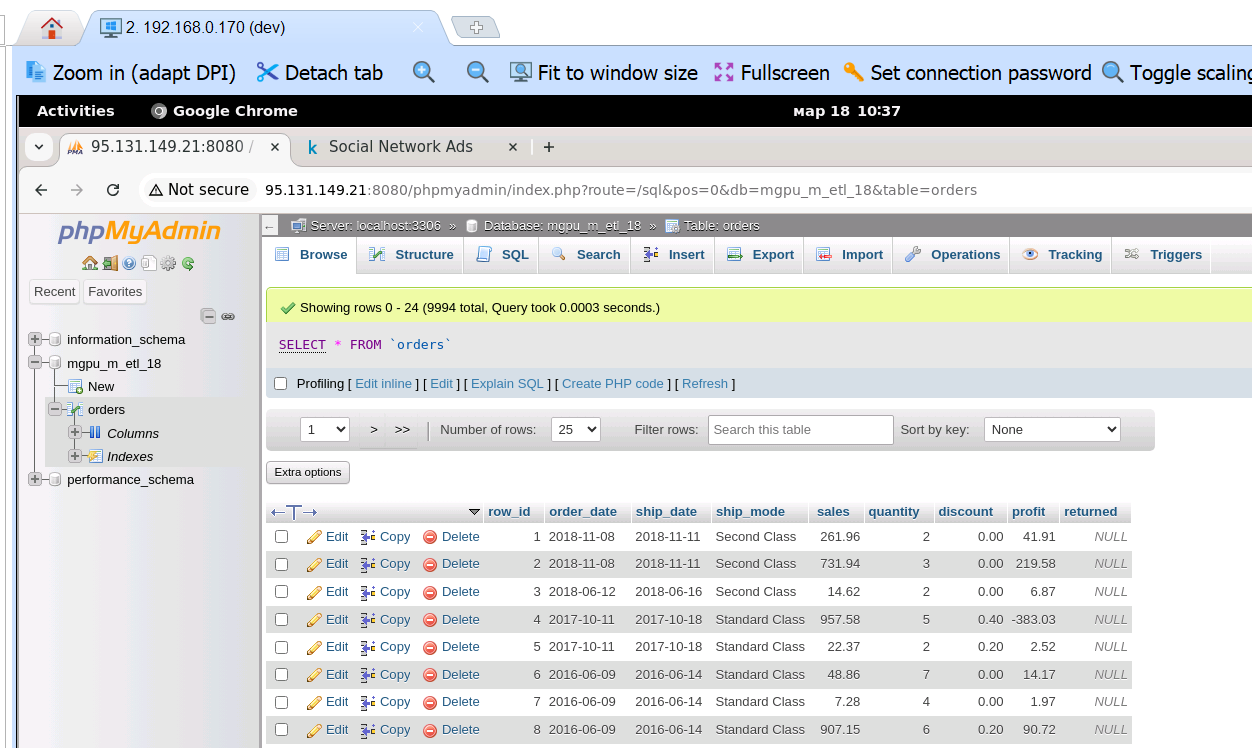
Москва

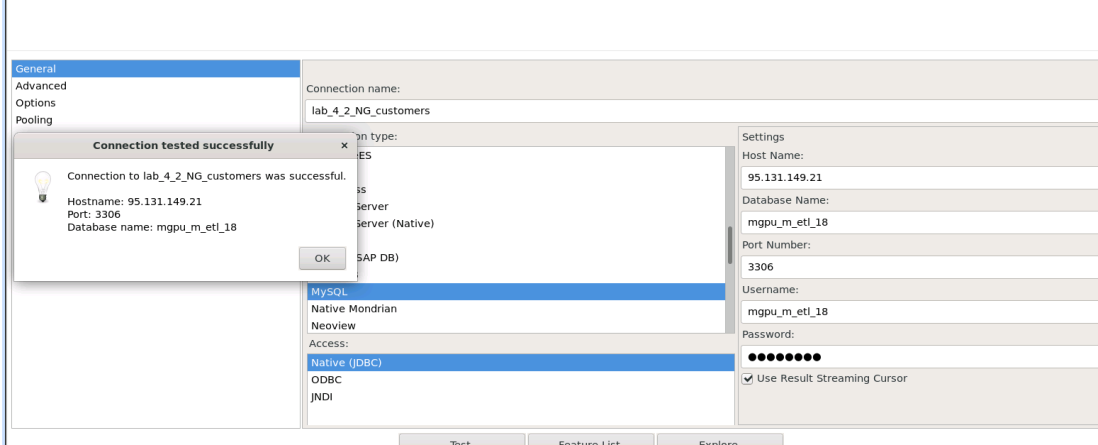
2024

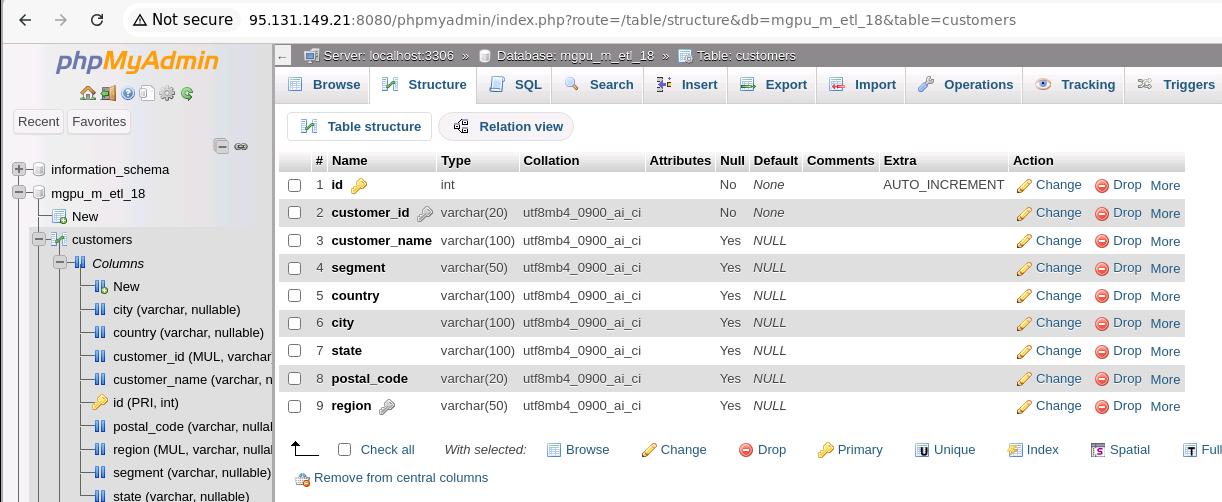
**Цель работы**: получить практические навыки создания ETL-процесса для загрузки данных из

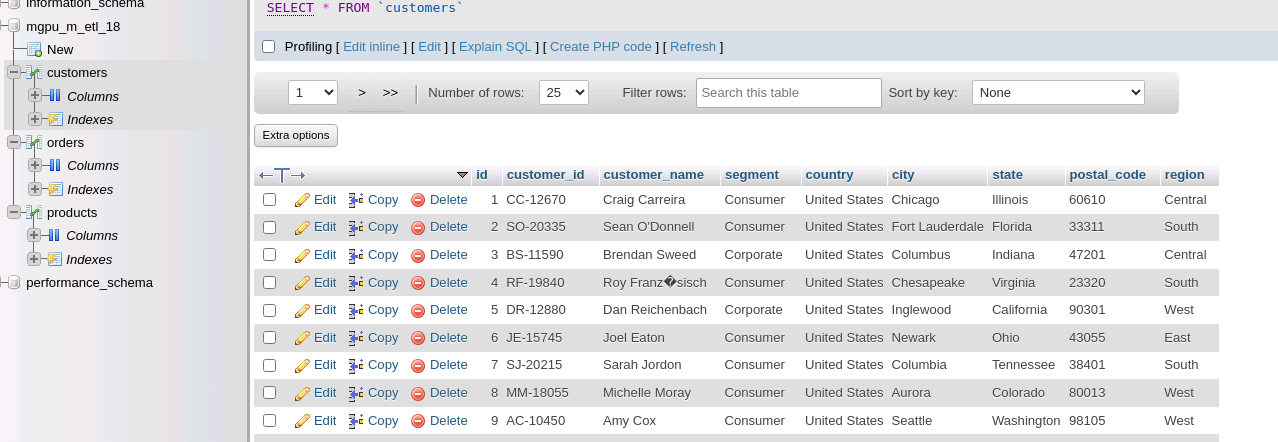
CSV-файла в базу данных MySQL с использованием Pentaho Data Integration.

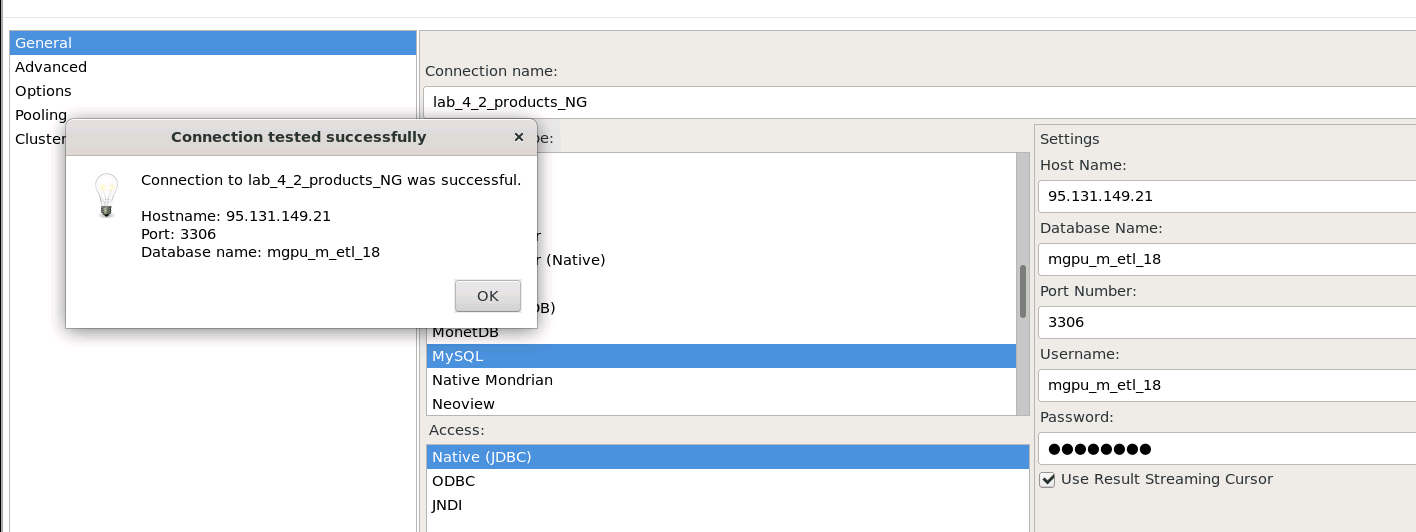


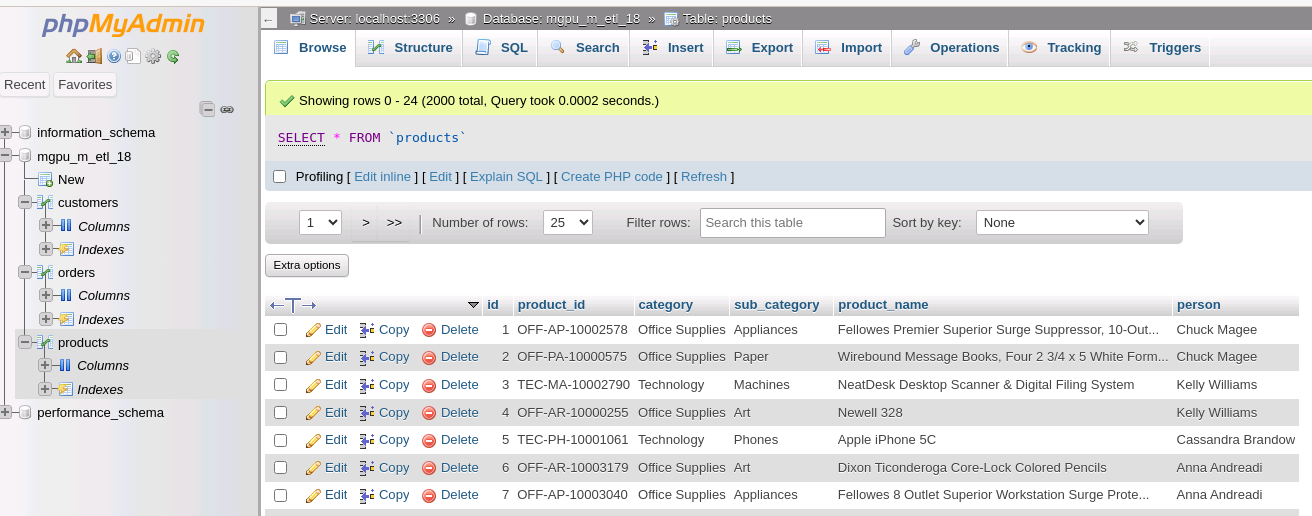




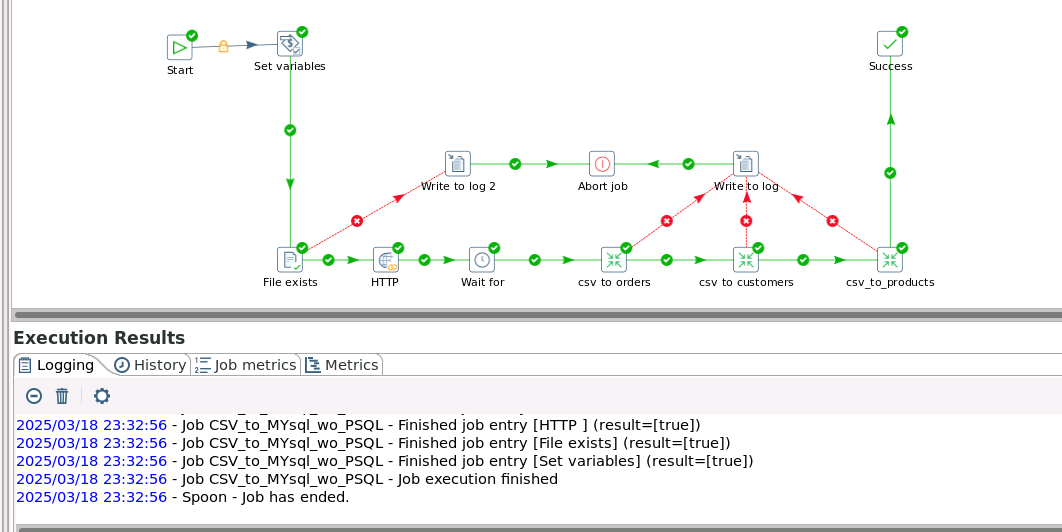


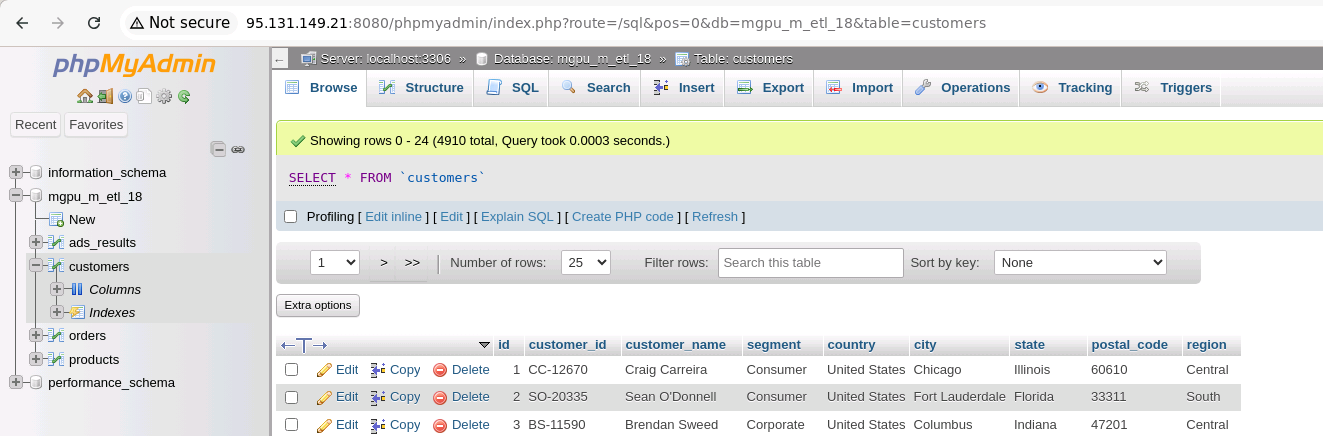


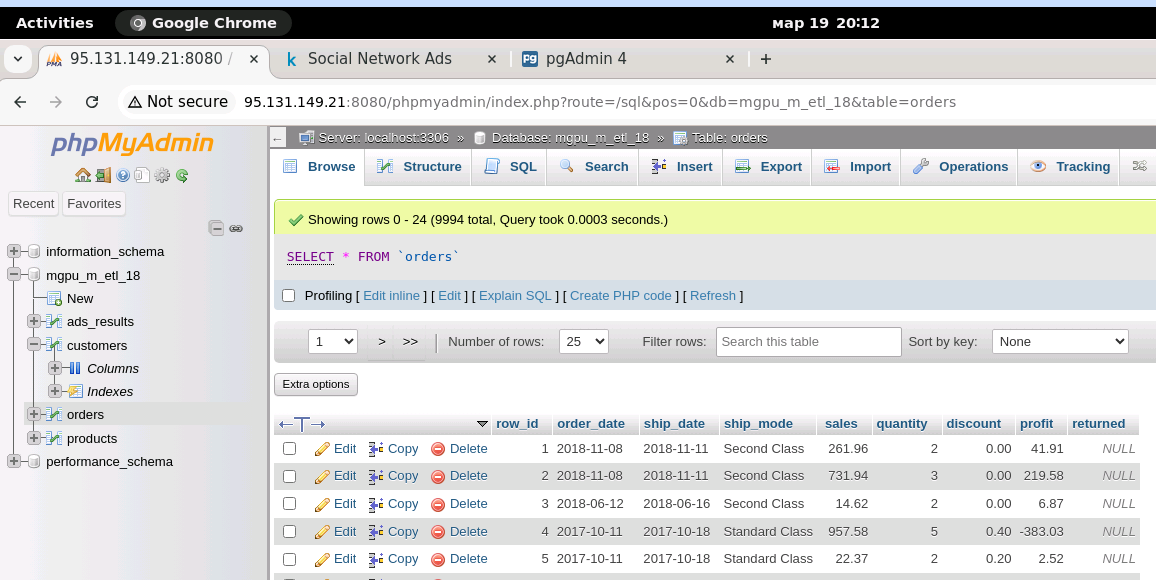


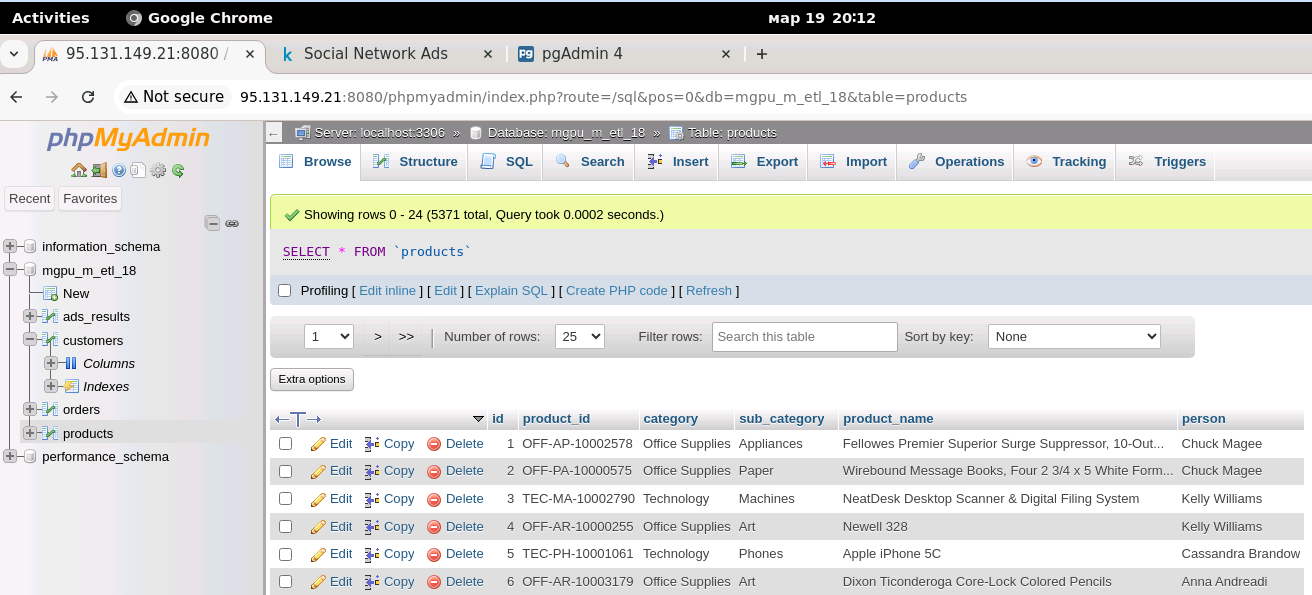


И вариант с использованием job

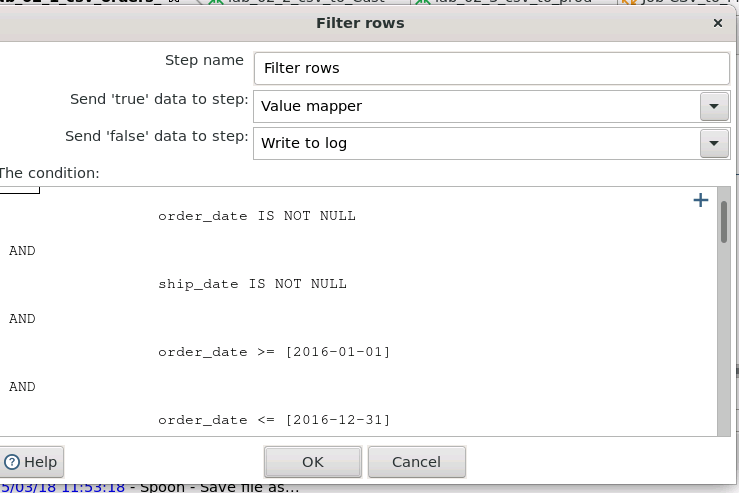


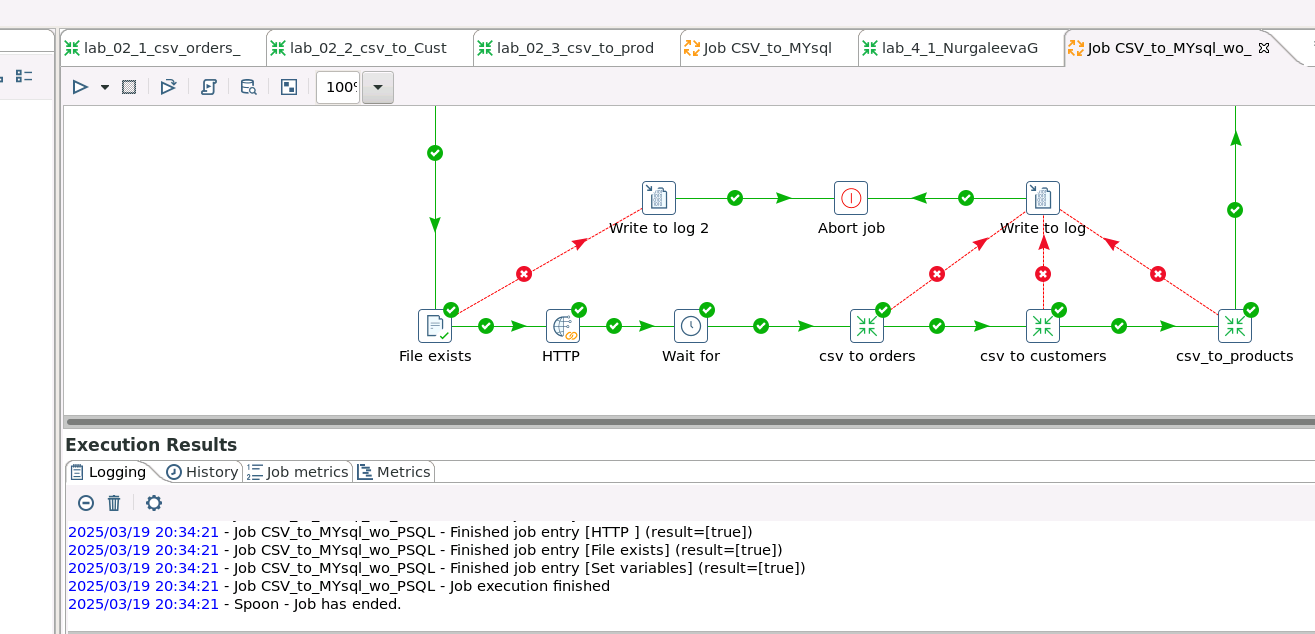


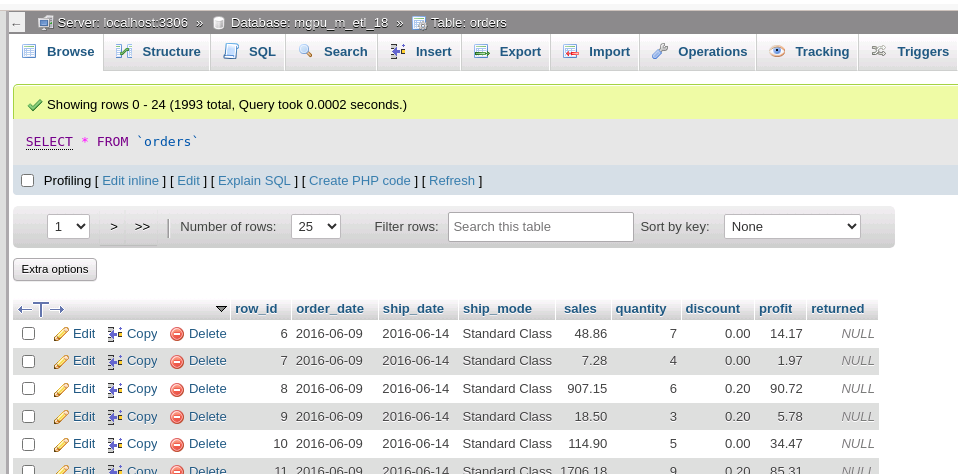




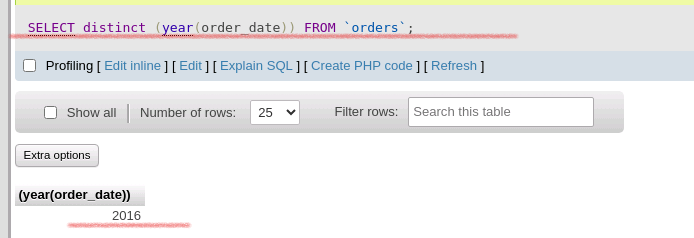
**Индивидуальное задание 18:** загрузить данные с фильтром по дате – только заказы 2016 года







**Проверка:**



**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.

1. **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.

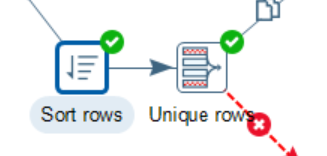
**- abrt-execution\_error** - прерывание выполнения при критических ошибках.

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

- можно удалить дубли используя группировку по всем столбцам



- через использование функции unique rows



1. **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.
4. **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**