

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

Кузьмина Дарья Юрьевна БД-241м

### **Практическая работа**

**Лабораторная работа 4.3. Интеграция данных из нескольких источников.**

### **Вариант 11**

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

Руководитель дисциплины:  
Босенко Т.М., доцент департамента  
информатики, управления и технологий,  
доктор экономических наук

Москва  
2025

## Содержание

Введение .....	2
Основная часть .....	2
Заключение.....	4

## Введение

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

### Задачи:

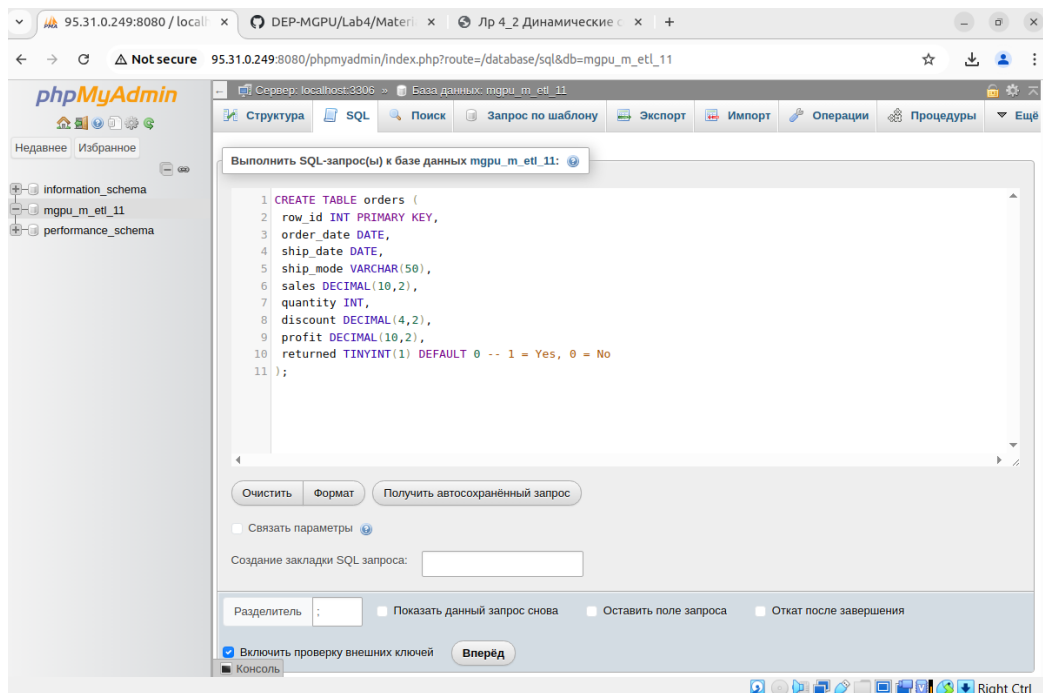
- Создать динамические подключения к различным источникам данных.
- Разработать процесс выявления и обработки дублирующихся записей.
- Реализовать механизм объединения данных в единое хранилище.
- Настроить обработку ошибок при выполнении трансформации.

## Основная часть

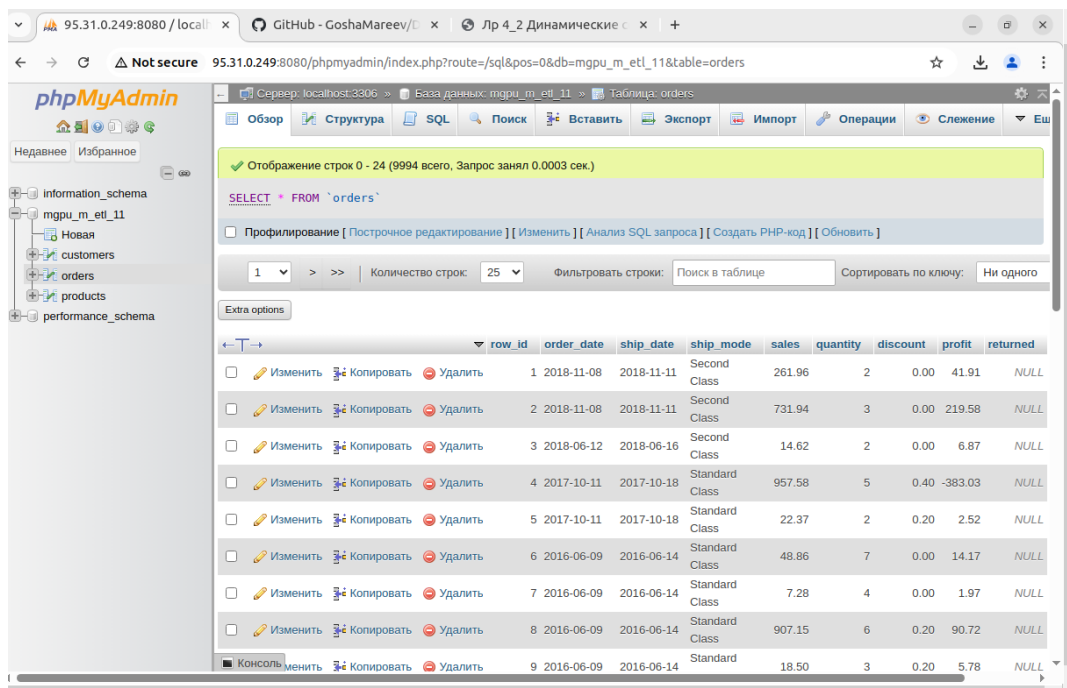
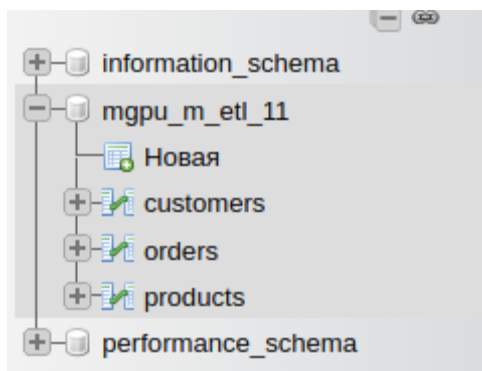
The screenshot displays the Pentaho Data Integration (Spoon) interface for a job named "lab\_02\_1\_csv\_orders". The main workspace shows a job flow diagram with the following steps: "CSV file input", "Select values", "Memory group by", "Filter rows", "Value mapper", and "table orders". A yellow box labeled "Поток Orders" is positioned above the "Filter rows" step. The left sidebar shows the "Transformations" tree with the current job selected. The top status bar indicates the job is running at 100%.

The "Execution Results" panel at the bottom provides a detailed log of the job execution:

- 2025/03/22 22:29:53 - CSV file input.0 - Header row skipped in file 'file:///home/dev/Downloads/lab\_etl/data\_for\_labs/lab\_csvexl\_to\_mysql/dat...
- 2025/03/22 22:29:54 - CSV file input.0 - Finished processing (I=9995, O=0, R=0, W=9994, U=0, E=0)
- 2025/03/22 22:29:55 - Select values.0 - Finished processing (I=0, O=0, R=9994, W=19988, U=0, E=0)
- 2025/03/22 22:29:55 - Dummy (do nothing).0 - Finished processing (I=0, O=0, R=9994, W=9994, U=0, E=0)
- 2025/03/22 22:29:56 - Memory group by.0 - Finished processing (I=0, O=0, R=9994, W=9994, U=0, E=0)
- 2025/03/22 22:29:56 - Filter rows.0 - Finished processing (I=0, O=0, R=9994, W=9994, U=0, E=0)
- 2025/03/22 22:29:56 - Value mapper.0 - Finished processing (I=0, O=0, R=9994, W=9994, U=0, E=0)



Делаем это с остальными таблицами



The screenshot shows the pgAdmin 4 web interface in a browser. The top navigation bar includes 'Файл', 'Объект', 'Инструменты', and 'Справка'. The main toolbar contains icons for file operations, query execution, and other database functions. The 'Запрос' (Query) tab is active, displaying the following SQL query:

```
1 SELECT * FROM public.employees
2 ORDER BY id ASC
```

The 'Data Output' tab shows the results of the query in a table format. The table has 9 columns: id, first\_name, last\_name, email, phone\_number, hire\_date, job\_title, and salary. The results are displayed in a grid with 10 rows. The status bar at the bottom indicates 'Total rows: 20' and 'Query complete 00:00:00.925'.

	id [PK] integer	first_name character varying (50)	last_name character varying (50)	email character varying (100)	phone_number character varying (20)	hire_date date	job_title character varying (50)	salary numeri
1	1	John	Doe	john.doe@example.com	1234567890	2015-06-15	Developer	
2	2	Jane	Smith	jane.smith@example.com	0987654321	2018-03-22	Manager	
3	3	Michael	Johnson	michael.johnson@example...	1112223333	2012-08-10	Analyst	
4	4	Emily	Davis	emily.davis@example.com	4445556666	2017-11-05	Designer	
5	5	David	Wilson	david.wilson@example.com	7778889999	2016-04-20	Developer	
6	6	Sarah	Brown	sarah.brown@example.com	2223334444	2019-09-12	Tester	
7	7	Chris	Lee	chris.lee@example.com	5556667777	2014-02-18	Analyst	
8	8	Jessica	Clark	jessica.clark@example.com	8889990000	2013-07-03	Manager	
9	9	Andrew	Garcia	andrew.garcia@example.com	3334445555	2011-12-25	Developer	
10	10	Linda	Martinez	linda.martinez@example.co...	6667778888	2010-05-14	Designer	

## Заключение

Целью работы было получение навыков интеграции, обработки и согласования данных из различных источников. Были изучены методы чтения данных из CSV, Excel и баз данных (MySQL, PostgreSQL), а также техники их очистки и согласования. Обработанные данные были сохранены в единое хранилище для дальнейшего анализа.

- Интеграция данных из нескольких источников требует тщательной подготовки и согласования форматов данных.
- Использование ETL-процессов позволяет автоматизировать обработку и объединение данных.
- Качество данных напрямую влияет на точность аналитических выводов, поэтому важна предварительная очистка и проверка данных.