

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

ДИСЦИПЛИНА:

Проектный практикум по разработке ETL-решений

Самостоятельная работа 1

Интеграция данных из разных источников (баз данных)

Выполнила: st_105

Москва

2025

Цель работы: разработка ETL-процесса для интеграции данных между PostgreSQL и MySQL с использованием Pentaho Data Integration.

Задачи:

- Создать исходные таблицы в PostgreSQL с различными наборами данных.
- Настроить целевые таблицы в MySQL для приема данных.
- Разработать процессы трансформации данных в Pentaho.
- Реализовать механизмы обработки ошибок и валидации данных.
- Создать представления для связанных данных.

ПОДГОТОВКА К РАБОТЕ:

Запуск контейнера pgadmin и postgresql (рисунок 1)

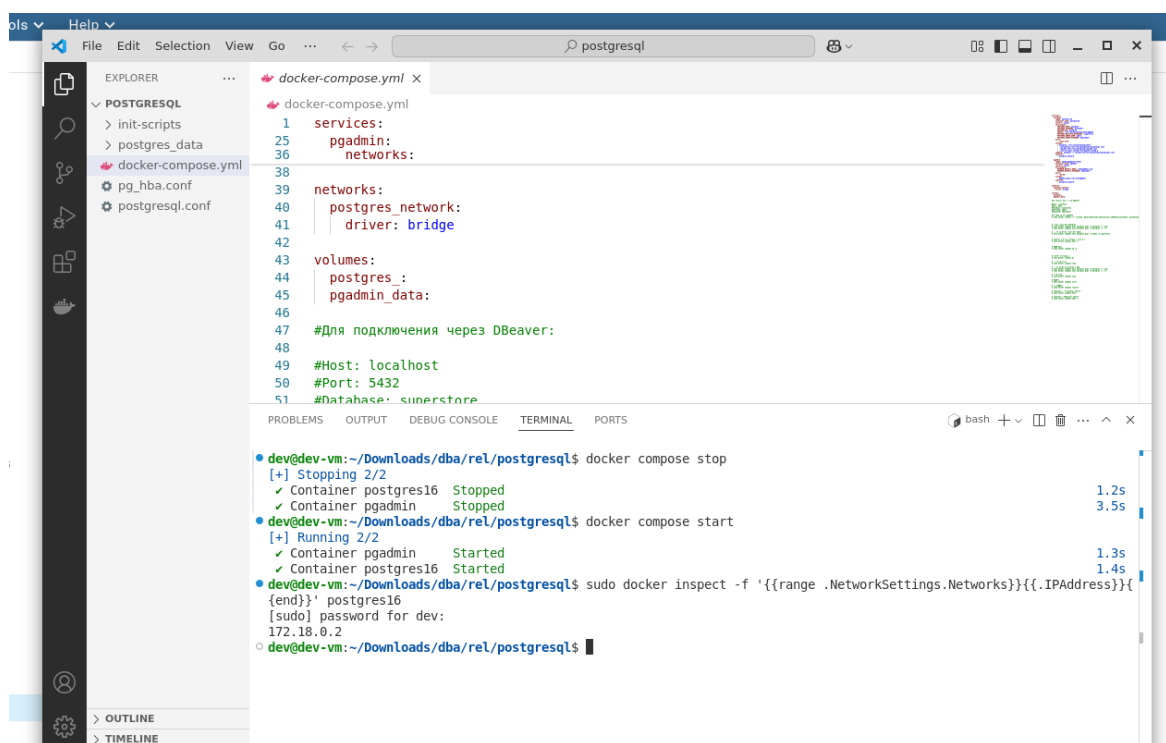


Рисунок 1 – запуск контейнеров для работы

Создание базы данных для выполнения задания st_105 (рисунок 2)

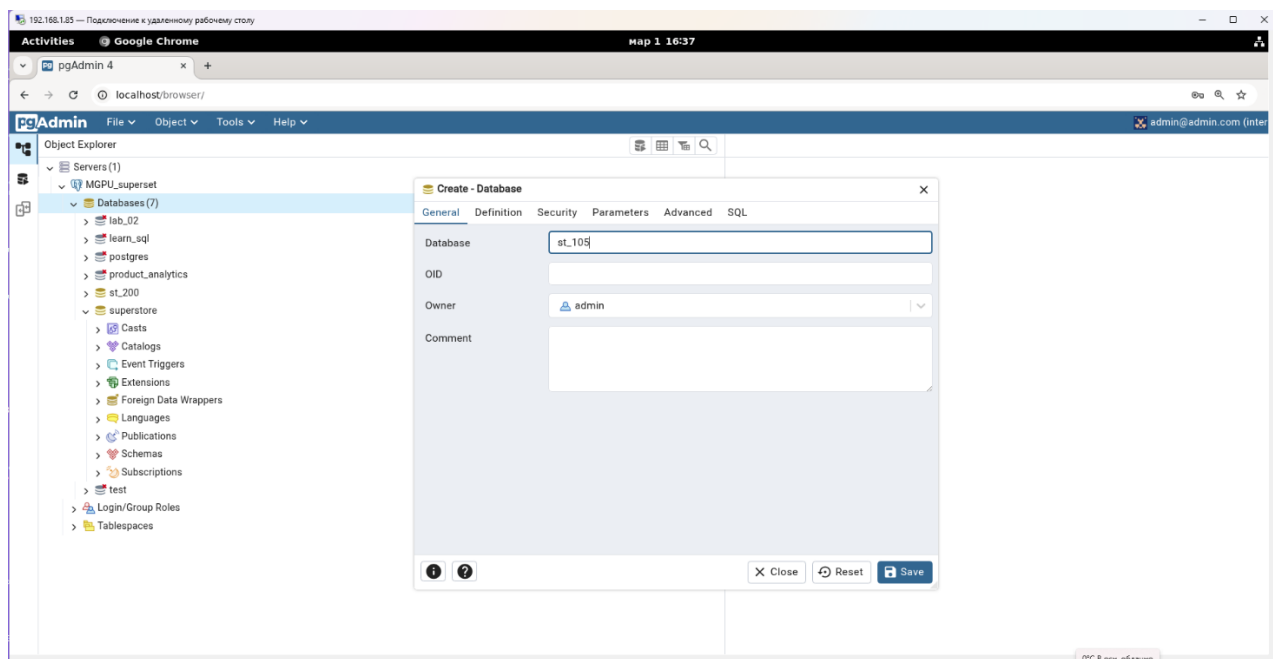


Рисунок 2 – создание бд

ХОД РАБОТЫ:

Вариант st_105

ID_студента	Задание 1 (PostgreSQL)	Задание 2 (MySQL)	Задание 3 (Pentaho)	Задание 4 (Pentaho)	Задание 5 (Pentaho)
st_105	Создать таблицу training (id, employee_id, course_name, start_date, status)	Создать таблицу training_progress с полем score	,Фильтр завершенных курсов	,Статистика по курсам	,Расчет успеваемости

Задание 1. Создание таблицы training в postgresql и наполнение таблицы сгенерированными данными (рисунок 3-4)

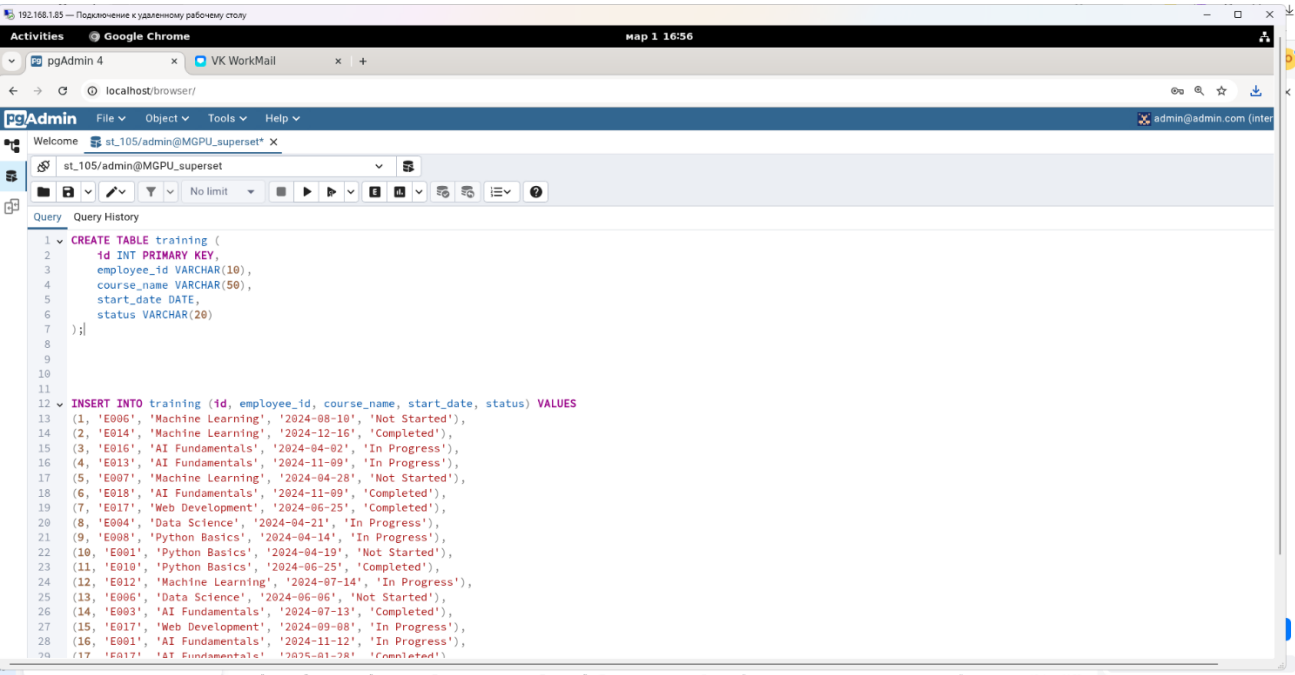


Рисунок 3 – ddl скрипт и insert скрипт

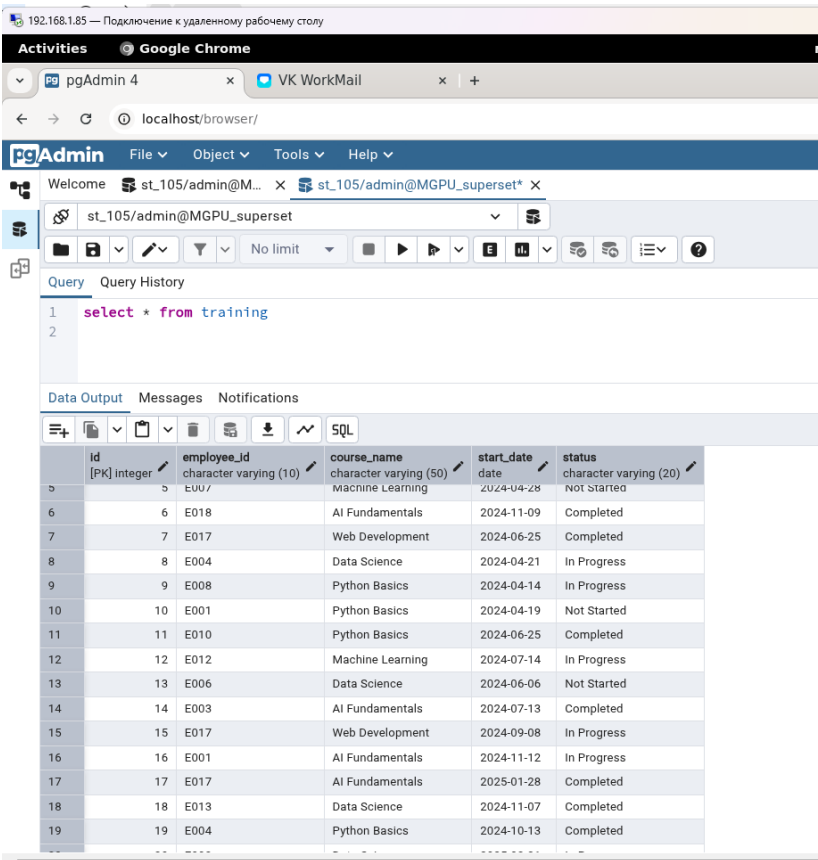


Рисунок 4 – проверка значений

Задание 2: Создание таблицы с полем score в Mysql.

Таблица состоит из трех полей: первичный ключ, внешний ключ от таблицы training, который ссылается на запись в таблицы где курс в статусе “in progress” или “Completed”, чтобы не было записей с успеваемостью, где участники не приступили к курсу, поле score – успеваемость от 0 до 100 (рисунок 5-7)

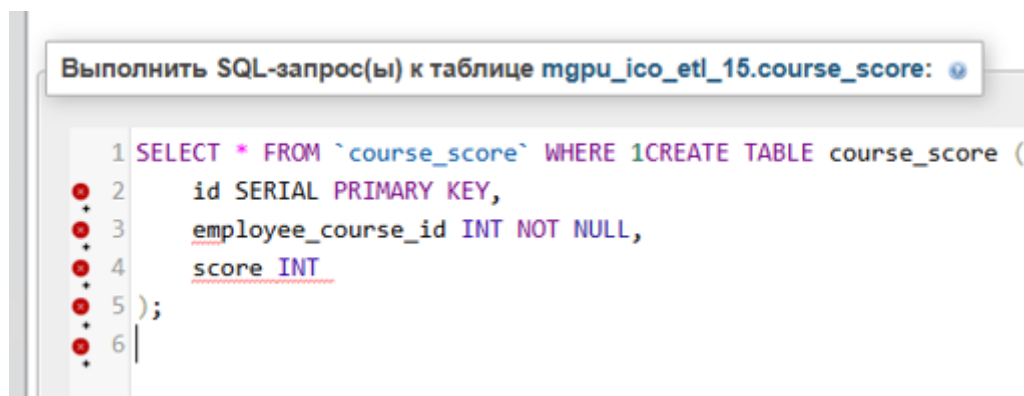


Рисунок 5 - ddl скрипт

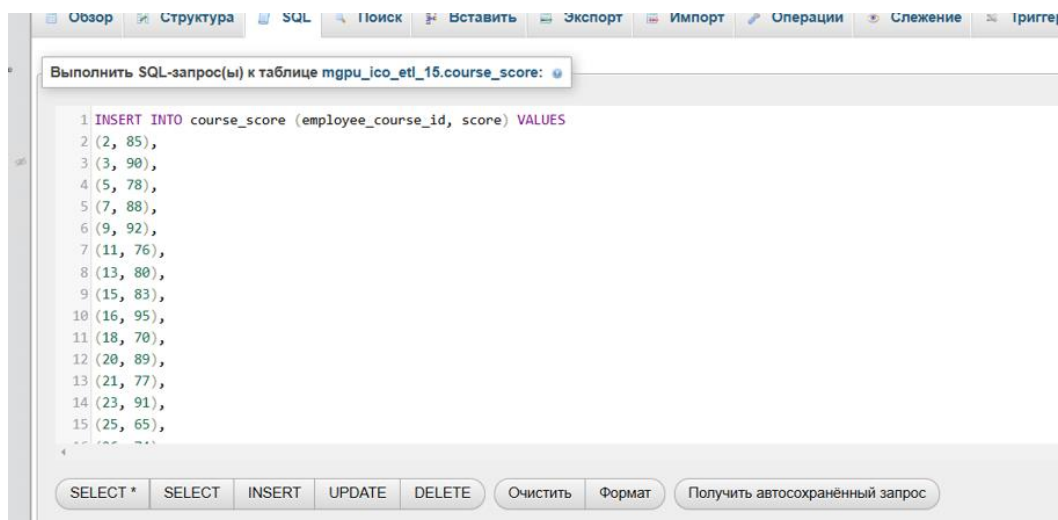


Рисунок 6 – вставка данных в таблицу

`SELECT * FROM `course_score``

☐ Профилирование [[Построчное редактирование](#)] [[Изменить](#)] [[Анализ SQL запроса](#)] [[Создать Р](#)]

1 > >> | ☐ Показать все | Количество строк: 25 | Фильтровать строки:

Extra options

	id	employee_course_id	score
<input type="checkbox"/> Изменить Копировать Удалить	1	2	85
<input type="checkbox"/> Изменить Копировать Удалить	2	3	90
<input type="checkbox"/> Изменить Копировать Удалить	3	5	78
<input type="checkbox"/> Изменить Копировать Удалить	4	7	88
<input type="checkbox"/> Изменить Копировать Удалить	5	9	92
<input type="checkbox"/> Изменить Копировать Удалить	6	11	76
<input type="checkbox"/> Изменить Копировать Удалить	7	13	80
<input type="checkbox"/> Изменить Копировать Удалить	8	15	83
<input type="checkbox"/> Изменить Копировать Удалить	9	16	95
<input type="checkbox"/> Изменить Копировать Удалить	10	18	70
<input type="checkbox"/> Изменить Копировать Удалить	11	20	89
<input type="checkbox"/> Изменить Копировать Удалить	12	21	77
<input type="checkbox"/> Изменить Копировать Удалить	13	23	91

Рисунок 7 – проверка данных в таблице Mysql

Задание 3: Фильтр завершенных курсов в Pentaho

Создаем таблицу в mysql для отчёта по курсам со статусом ‘completed’

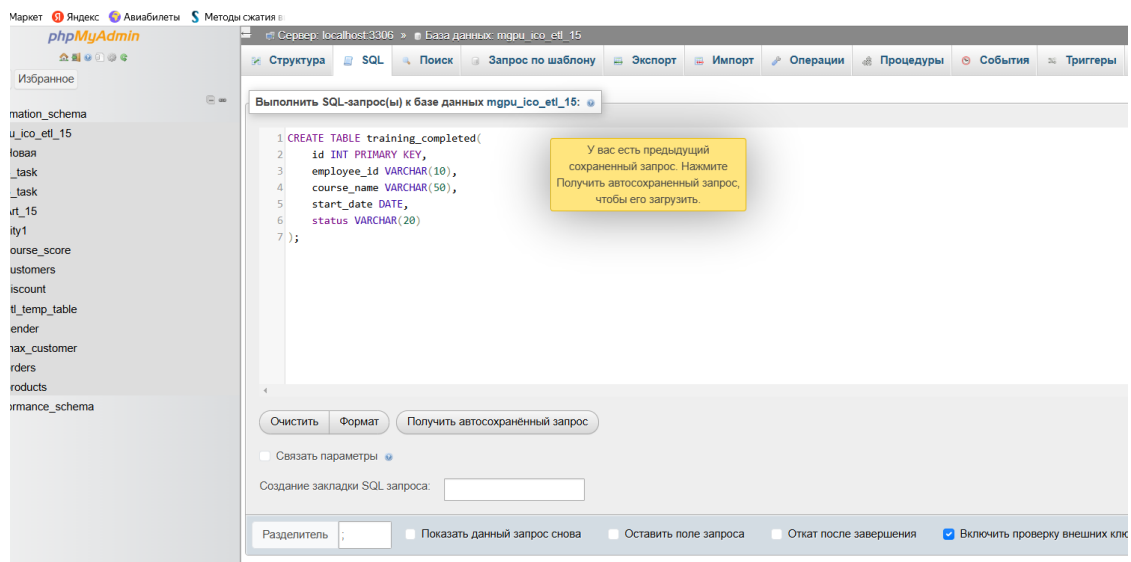


Рисунок 8 – ddl скрипт таблицы training_completed

Создаем трансформацию:

Настраиваем узлы:

— PostgreSQL

- Подключение postgresql (рисунок 9)
- Выбор полей из таблицы training (рисунок 10)
- **Настройки фильтра для таблицы training по значению статуса курсов «completed» (рисунок 11-12)**
- Отчёт в rhrtuadmin (рисунок 13)

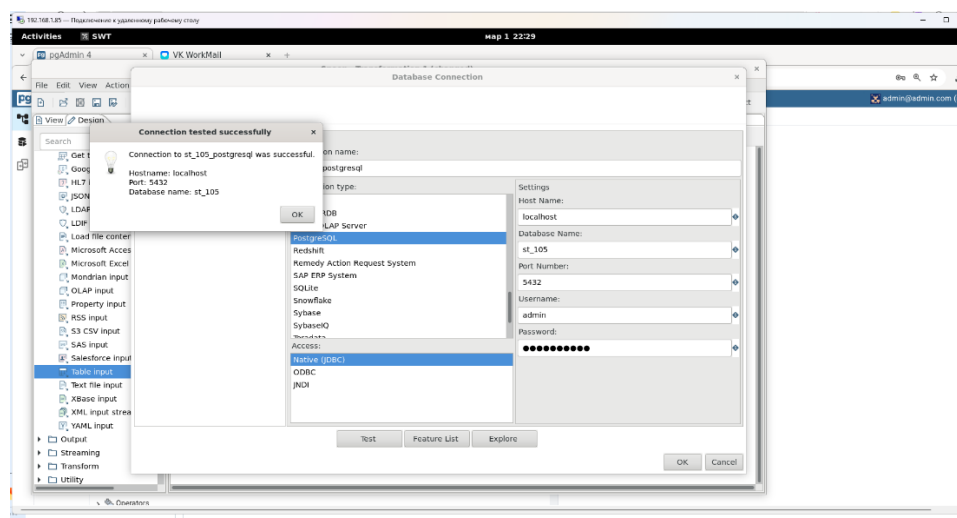


Рисунок 9 – проверка подключения к бд в pgadmin st_105

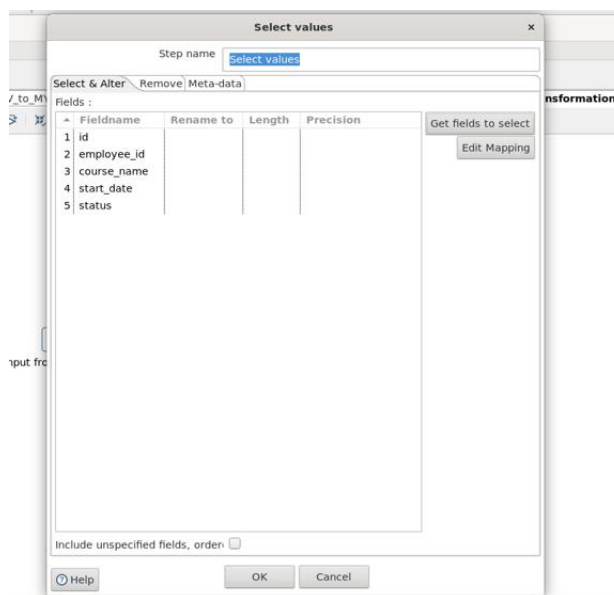


Рисунок 10 –выбор полей из таблицы training postgresql

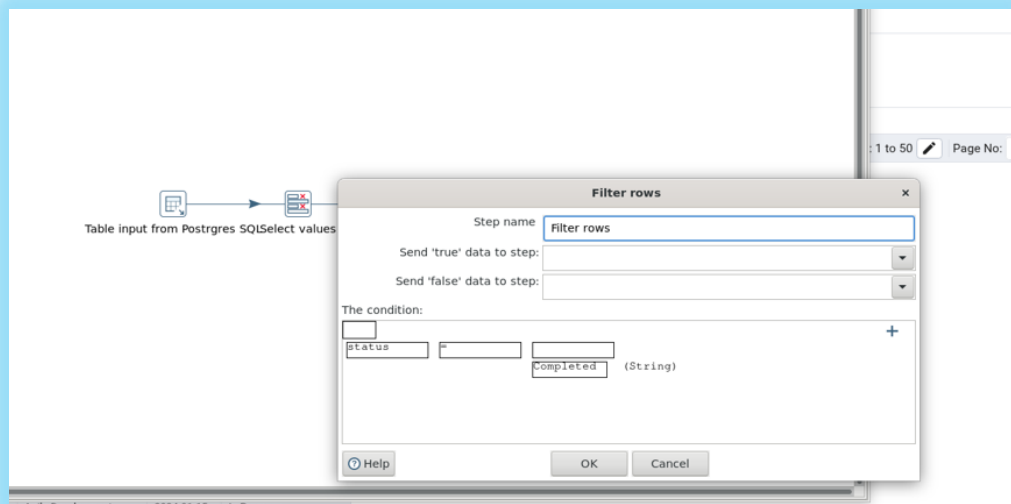


Рисунок 11 – настройка фильтра строк по значению completed

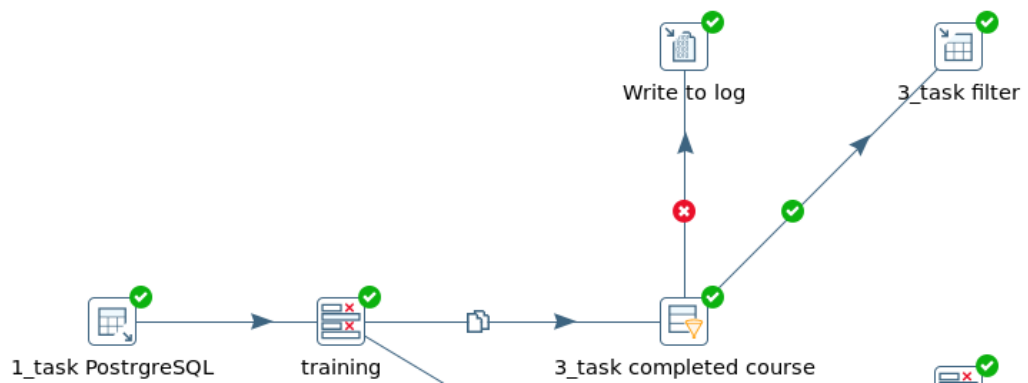


Рисунок 12 – трансформация на этапе задания 3

Отображение строк 0 - 14 (15 всего, Запрос занял 0.0001 сек.)

SELECT * FROM `training_completed`

Профилирование [Построчное редактирование] [Изменить] [Анализ SQL запроса] [Создать PHP-код] [Обновить]

Показать все | Количество строк: 25 | Фильтровать строки: Поиск в таблице | Сортировать по ключу: Ни од

Extra options

				id	employee_id	course_name	start_date	status
<input type="checkbox"/>	Изменить	Копировать	Удалить	2	E007	Data Science Basics	2024-01-10	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	7	E006	Agile Development	2024-03-01	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	9	E009	Big Data Processing	2024-01-25	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	13	E001	Cloud Computing	2024-02-28	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	16	E004	SQL Fundamentals	2024-03-05	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	20	E008	Artificial Intelligence	2023-11-30	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	23	E011	Project Management	2024-01-28	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	25	E001	Big Data Processing	2023-12-08	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	29	E005	SQL Fundamentals	2023-12-13	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	33	E009	Project Management	2024-02-18	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	36	E012	Data Science Basics	2024-01-20	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	40	E004	Big Data Processing	2024-02-22	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	43	E007	Machine Learning	2024-01-30	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	46	E010	Cybersecurity	2024-01-26	Completed
<input type="checkbox"/>	Изменить	Копировать	Удалить	50	E002	Big Data Processing	2024-01-29	Completed

Отметить все | С отмеченными: Изменить | Копировать | Удалить | Экспорт

Рисунок 13- итоговый отчёт по 3 заданию

Вывод по 3 заданию: Фильтр вывел только те курсы, которые находятся в статусе «завершен» то есть 15 строк из 50.

— Mysql

- Подключение Mysql (рисунок 14)
- Выбор полей из таблицы course_score (рисунок 15)

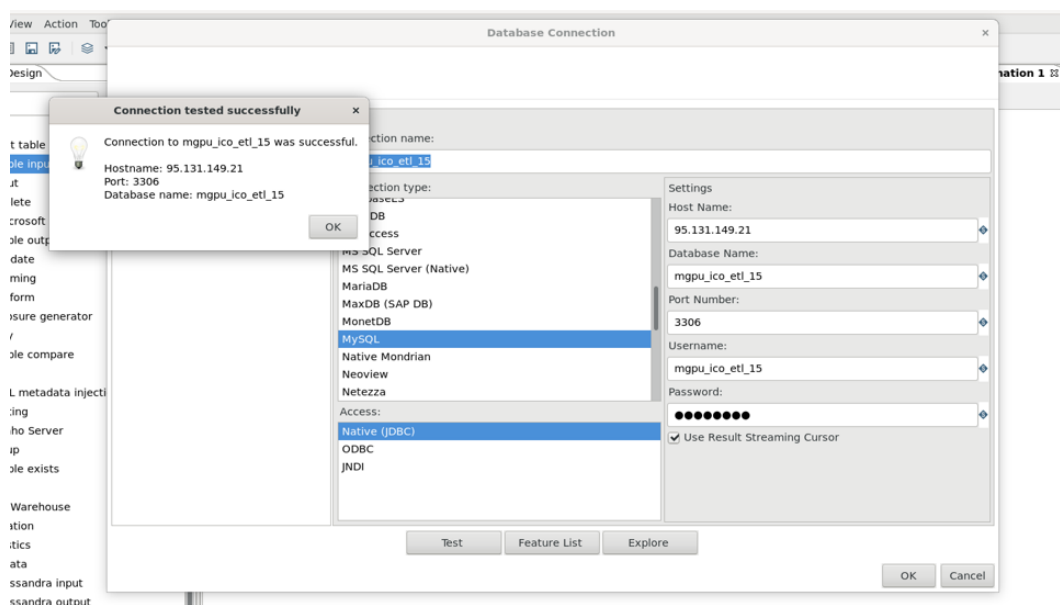


Рисунок 14 – настройка подключения mysql

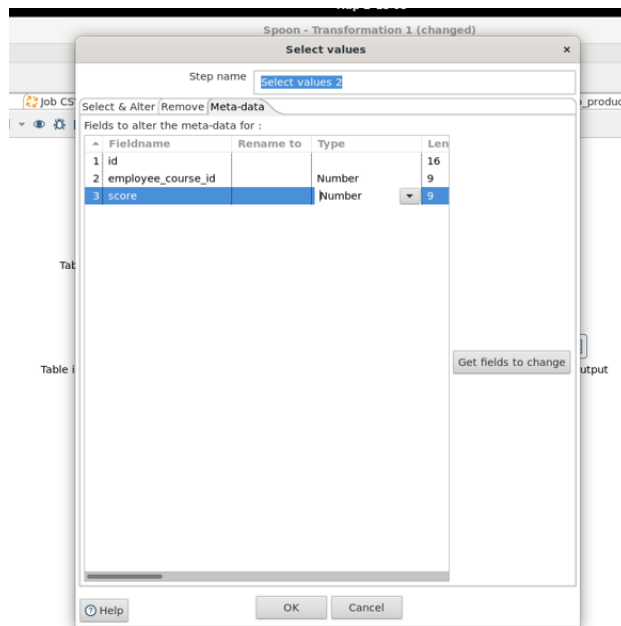


Рисунок 15 – выбор полей из таблицы

Задание 4: Статистика по курсам

Подключение из двух таблиц выполнено, далее настраиваем узел memory group by для составления отчёта по статистике, где будут следующие поля: course_name, status, count_student

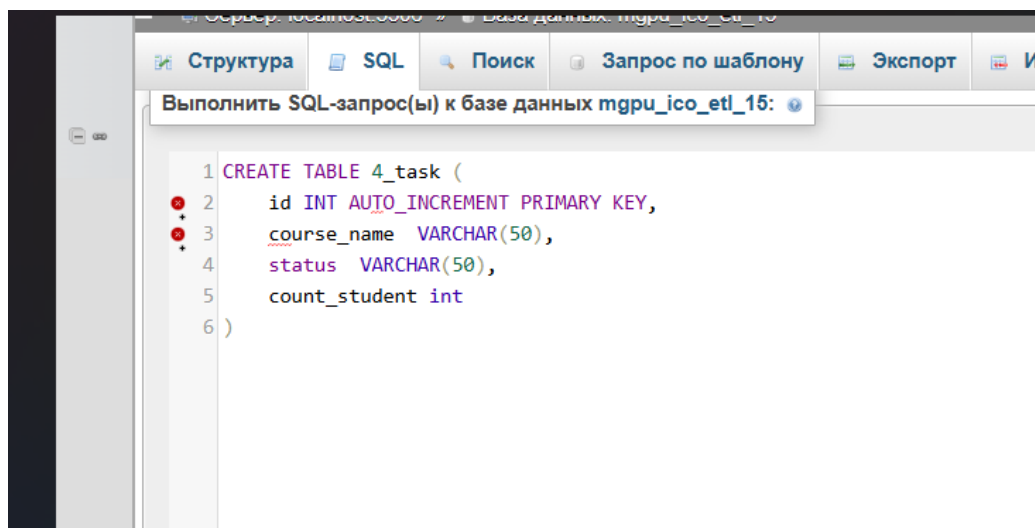


Рисунок 15 – ddl скрипт таблицы 4_task

Настройка узла группировки (рисунок 16)

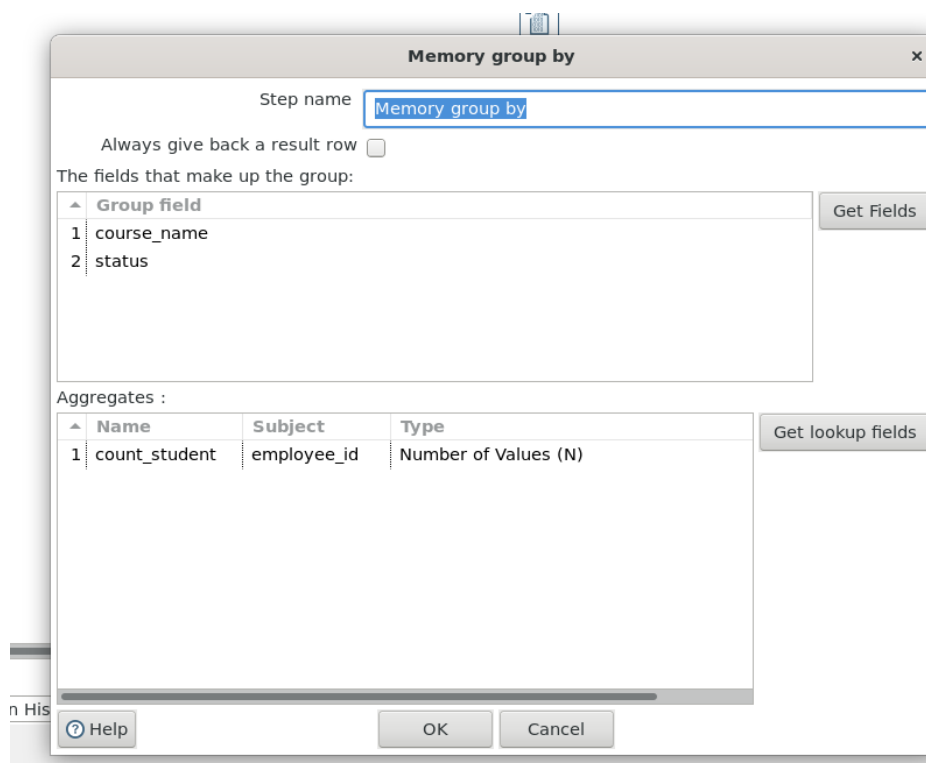


Рисунок 16 – группировка

Extra options

				id	course_name	status	count_student
<input type="checkbox"/>	Изменить	Копировать	Удалить	1	Agile Development	Completed	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	22	Agile Development	In Progress	3
<input type="checkbox"/>	Изменить	Копировать	Удалить	16	Artificial Intelligence	Completed	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	25	Artificial Intelligence	In Progress	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	6	Artificial Intelligence	Not Started	3
<input type="checkbox"/>	Изменить	Копировать	Удалить	21	Big Data Processing	Cancelled	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	7	Big Data Processing	Completed	4
<input type="checkbox"/>	Изменить	Копировать	Удалить	9	Cloud Computing	Cancelled	2
<input type="checkbox"/>	Изменить	Копировать	Удалить	11	Cloud Computing	Completed	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	10	Cloud Computing	In Progress	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	23	Cybersecurity	Completed	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	13	Cybersecurity	In Progress	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	20	Cybersecurity	Not Started	2
<input type="checkbox"/>	Изменить	Копировать	Удалить	26	Data Science Basics	Cancelled	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	19	Data Science Basics	Completed	2
<input type="checkbox"/>	Изменить	Копировать	Удалить	2	DevOps	Cancelled	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	30	DevOps	In Progress	2
<input type="checkbox"/>	Изменить	Копировать	Удалить	8	DevOps	Not Started	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	27	Machine Learning	Completed	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	14	Machine Learning	In Progress	1
<input type="checkbox"/>	Изменить	Копировать	Удалить	17	Machine Learning	Not Started	2
<input type="checkbox"/>	Изменить	Копировать	Удалить	28	Project Management	Cancelled	2

Рисунок 17 – итоговые данные в phrmyadmin

Вывод по заданию 4: итоговый отчёт показывает количество студентов в разрезе курсов и их статусов, например: Курс по DevOps находится в статусе «в процессе» у 2 человек, «отменен» у одного человека и «не начат» у одного человека

5 задание: Расчет успеваемости

Для отчёта успеваемости нужно объединить две таблицы training и course_score. Создаем таблицу в phpmyadmin с полями: employee_id, course_name, score (рисунок 18)

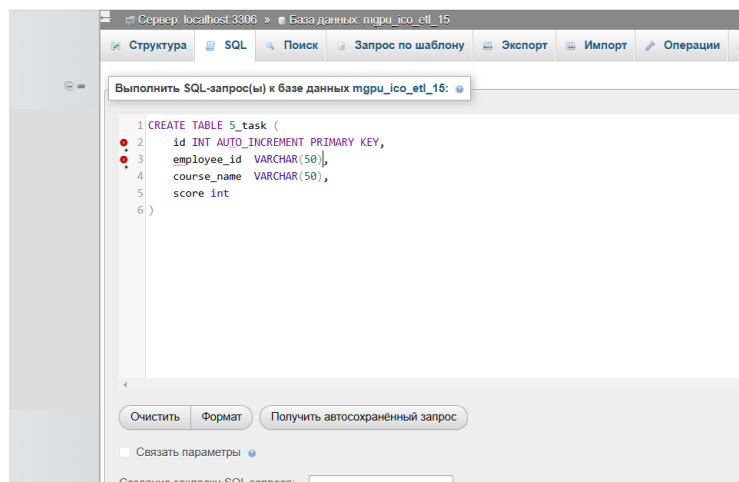


Рисунок 18 – ddl скрипт для 5_task

Настраиваем узел merge join по полю employee_course_id в таблице course_score и полю id из таблицы training (рисунок 19)

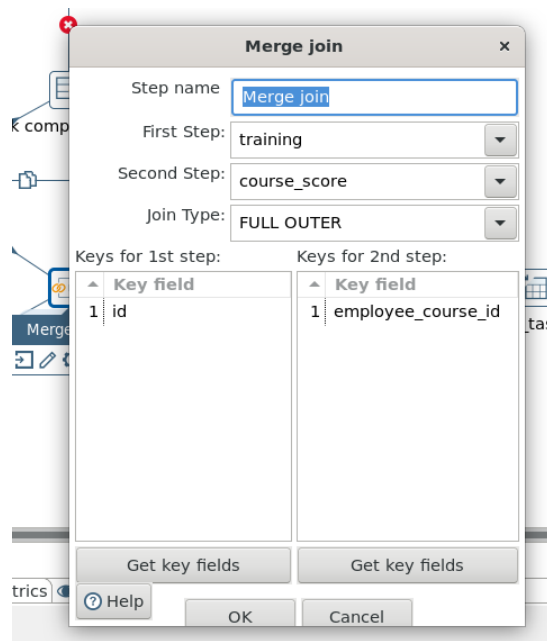


Рисунок 19 – настройка узла объединения

Так как успеваемость есть только в записях, где статус курса «начат» или «в процессе», следовательно, настраиваем узел фильтрации, который будет отсекал пустые строки из объединенной таблицы (рисунок 20)

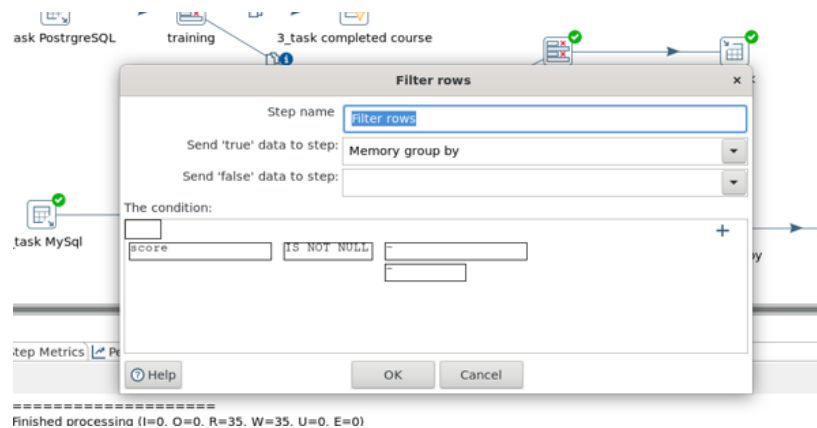


Рисунок 20 – фильтр пустых строк

Далее выбираем нужные столбцы с помощью select values (рисунок 21)

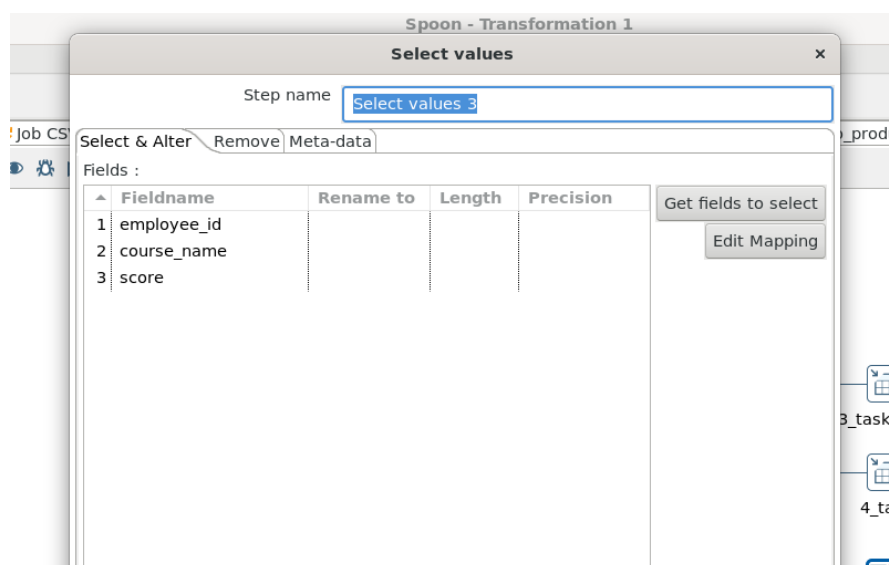


Рисунок 21 – настройка выбора полей

	id	employee_id	course_name	score
Удалить	26	E001	Artificial Intelligence	73
Удалить	14	E001	Big Data Processing	65
Удалить	7	E001	Cloud Computing	80
Удалить	2	E001	Python for Data Analysis	90
Удалить	15	E002	Agile Development	74
Удалить	27	E002	Big Data Processing	85
Удалить	3	E002	Cloud Computing	78
Удалить	8	E003	Agile Development	83
Удалить	21	E003	DevOps	88
Удалить	16	E004	Artificial Intelligence	86
Удалить	22	E004	Big Data Processing	90
Удалить	9	E004	SQL Fundamentals	95
Удалить	4	E006	Agile Development	88
Удалить	17	E006	Machine Learning	93
Удалить	10	E006	Python for Data Analysis	70
Удалить	1	E007	Data Science Basics	85
Удалить	23	E007	Machine Learning	81
Удалить	24	E008	Agile Development	87
Удалить	11	E008	Artificial Intelligence	89
Удалить	5	E009	Big Data Processing	92
Удалить	12	E009	Cybersecurity	77

Рисунок 22 – итоговый отчёт по заданию 5

Вывод по 5 заданию: Расчёт успеваемости показывает конкретного обучающегося, наименование курса и успеваемость по курсу. Например: E001 имеет успеваемость по курсу big data.. 65 и также имеет успеваемость 90 на курсе Python..

Трансформация со всеми заданиями (рисунок 23)

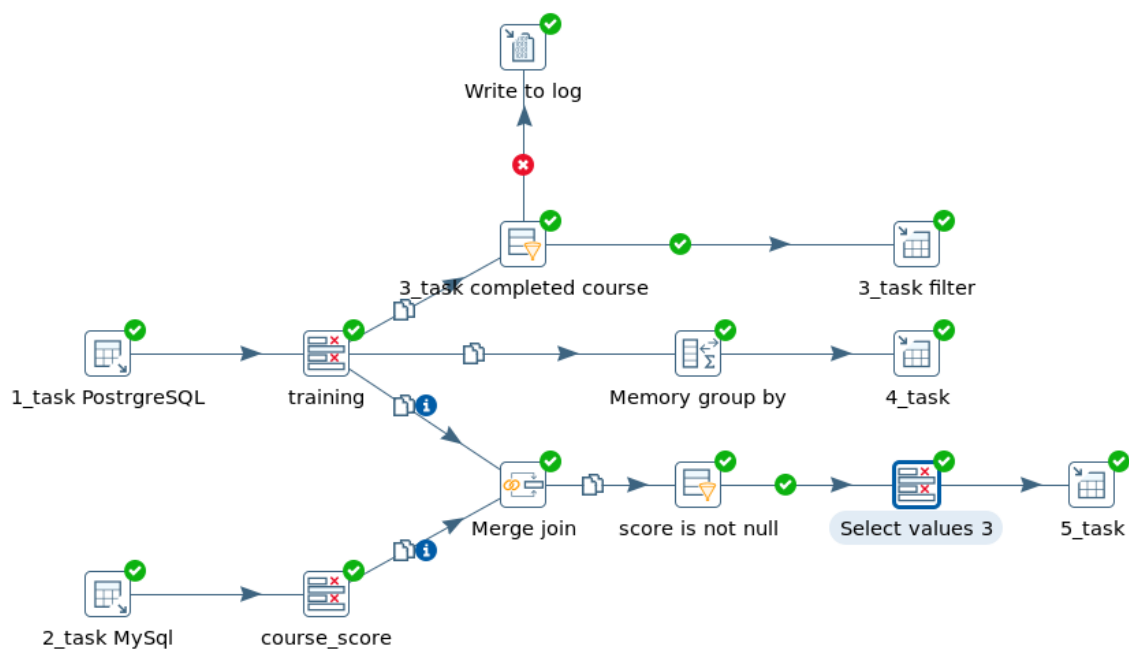


Рисунок 23 – общая трансформация

Вывод по работе: В ходе работы были импортированы данные из двух разных подключений mysql и postgresql. Выполнены задания и составлены отчёты.