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)

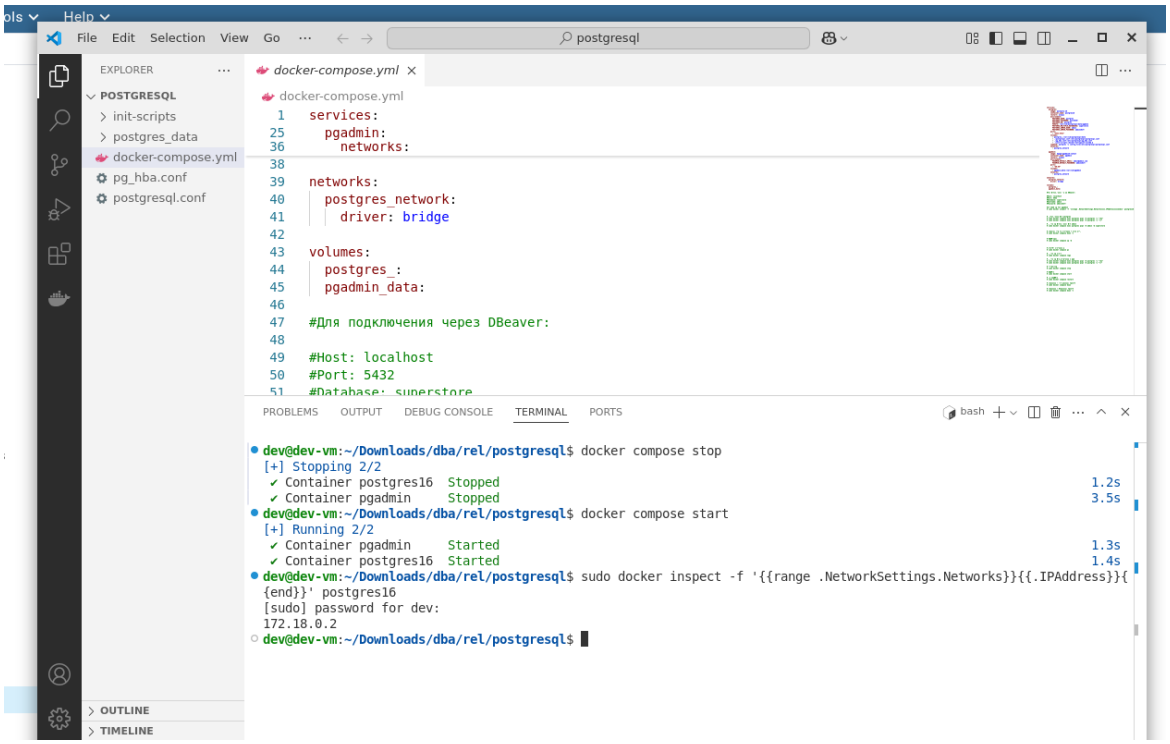

Рисунок 1 – запуск контейнеров для работы

Создание базы данных для выполнения задания st_105 (рисунок 2)

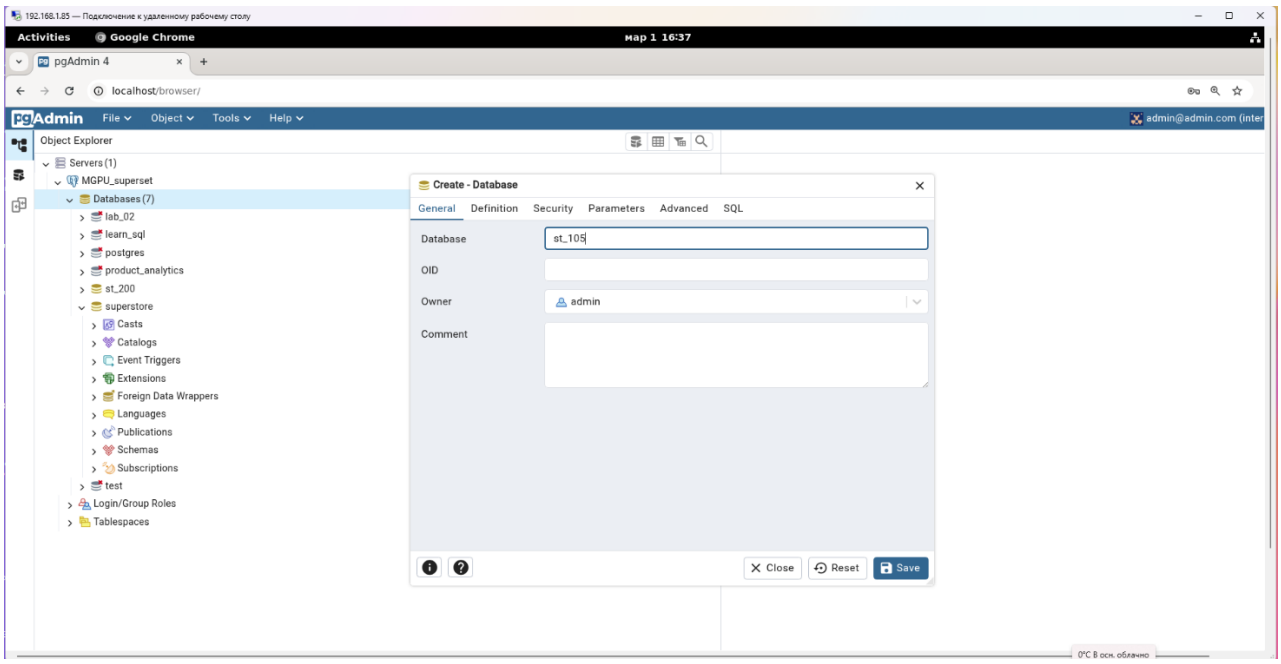

Рисунок 2 – создание бд

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)

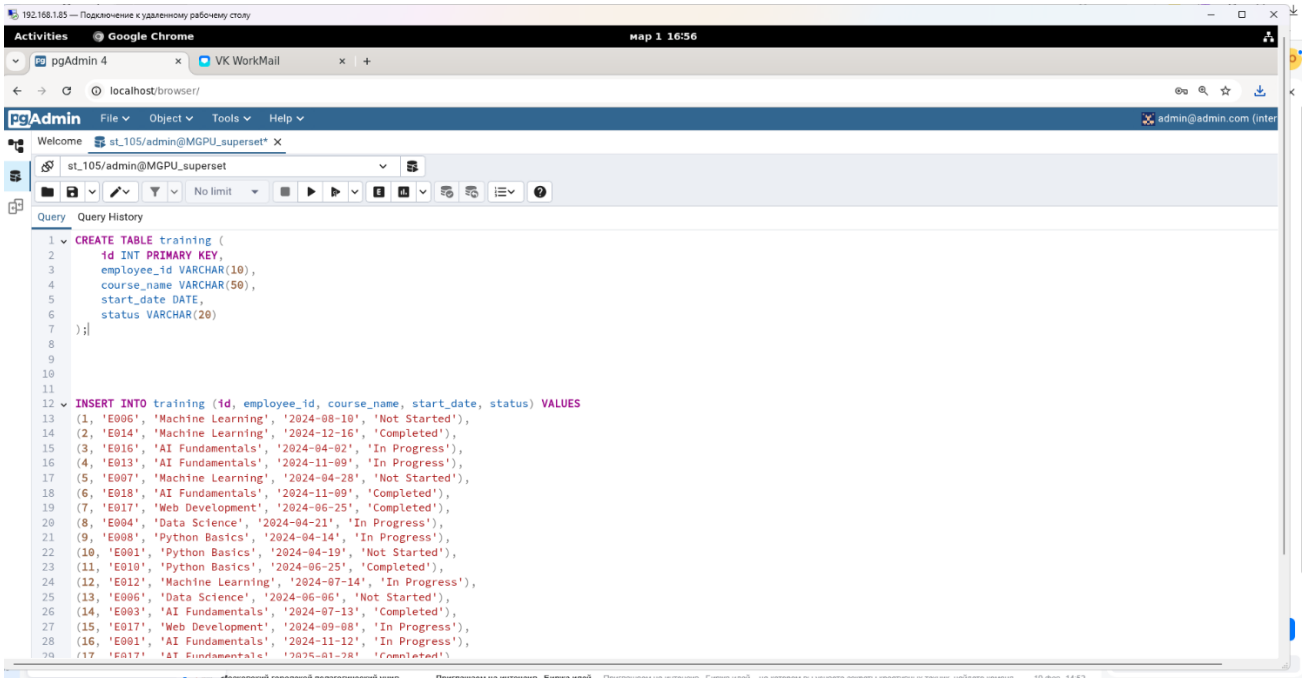

Рисунок 3 – ddl скрипт и insert скрипт

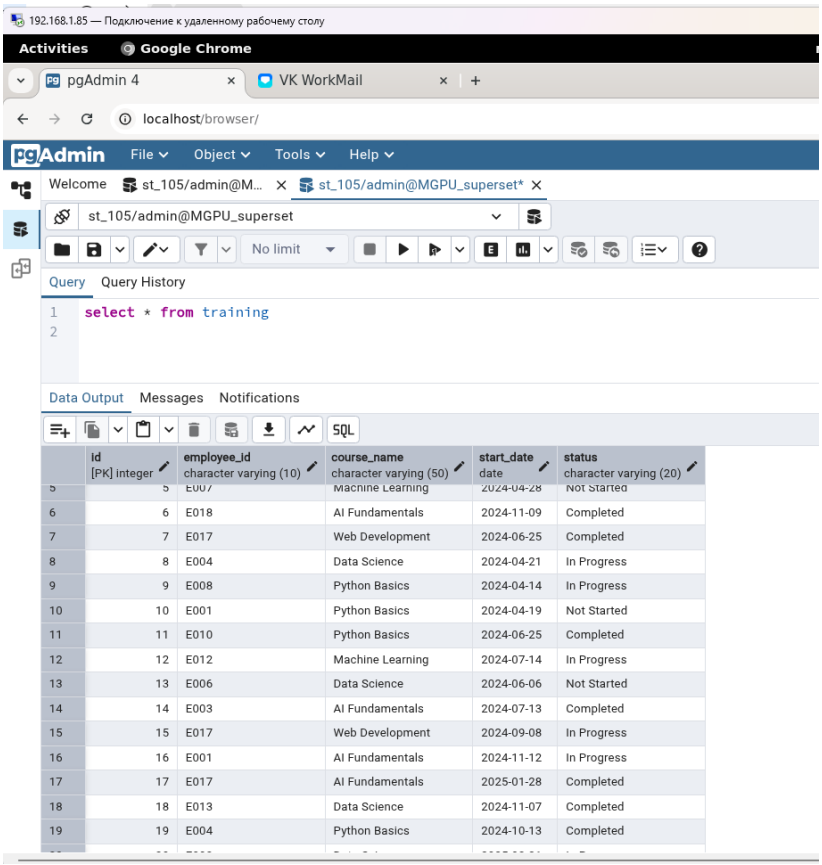

Рисунок 4 – проверка значений

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)

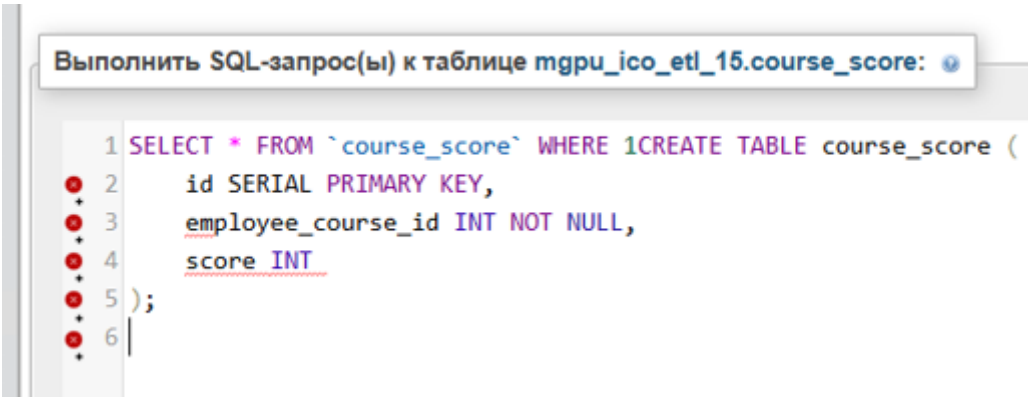

Рисунок 5 - ddl скрипт

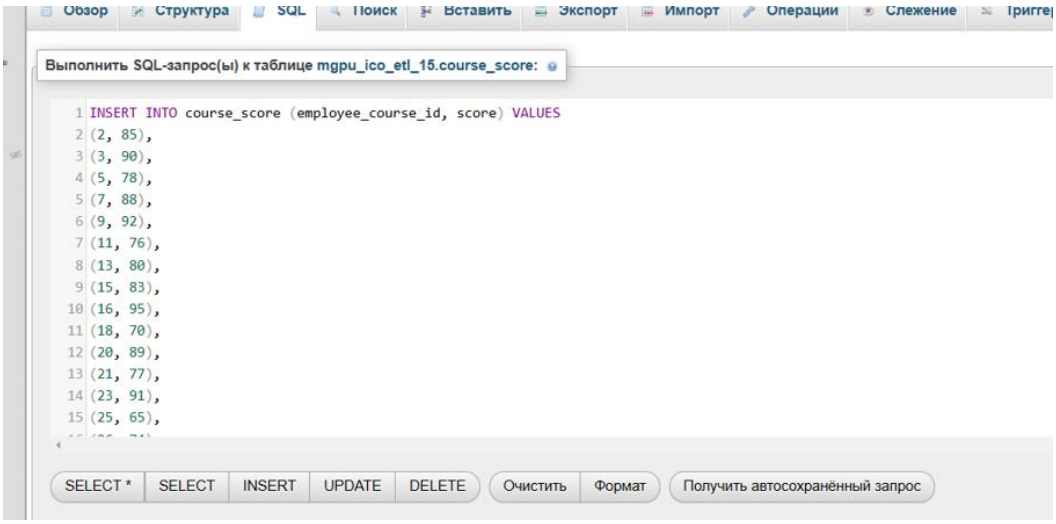

Рисунок 6 – вставка данных в таблицу

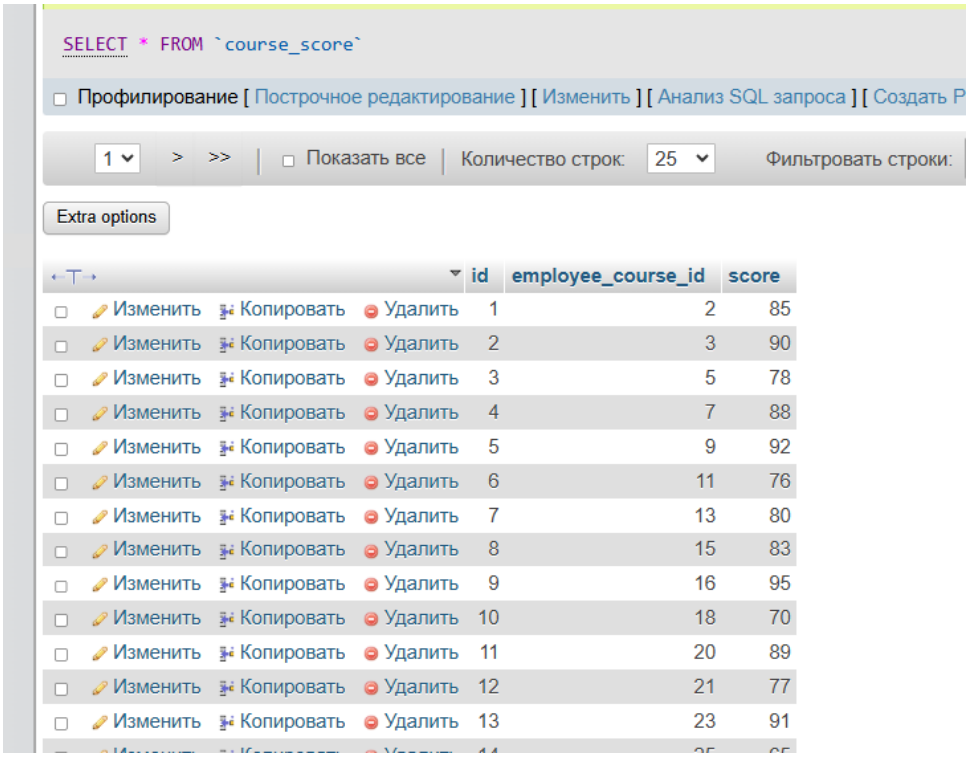

Рисунок 7 – проверка данных в таблице Mysql

Задание 3: Фильтр завершенных курсов в Pentaho

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

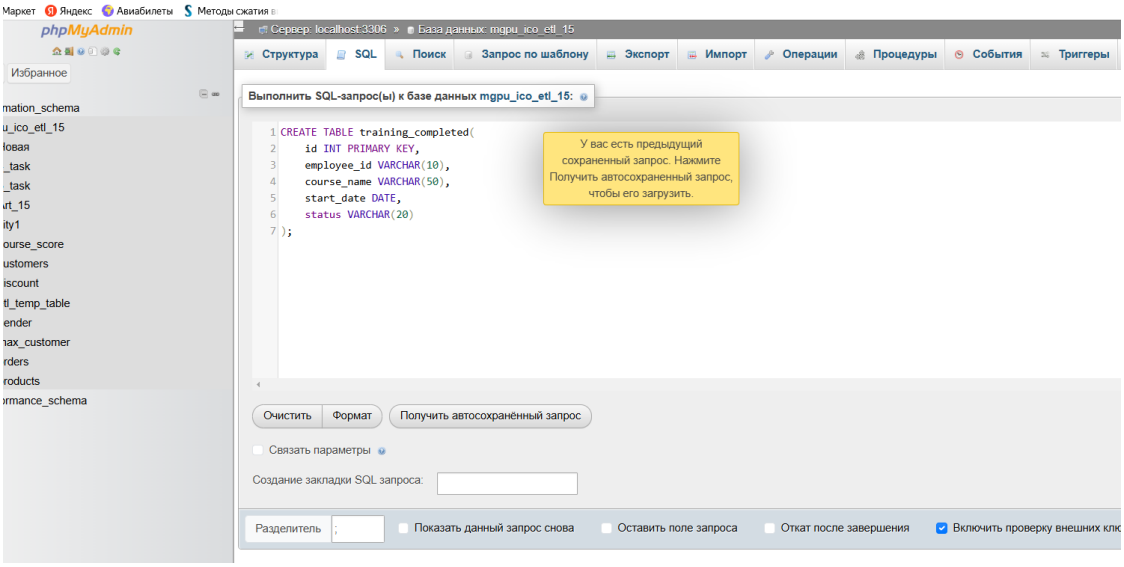

Рисунок 8 – ddl скрипт таблицы training_completed

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 phpmyadmin (рисунок 13)

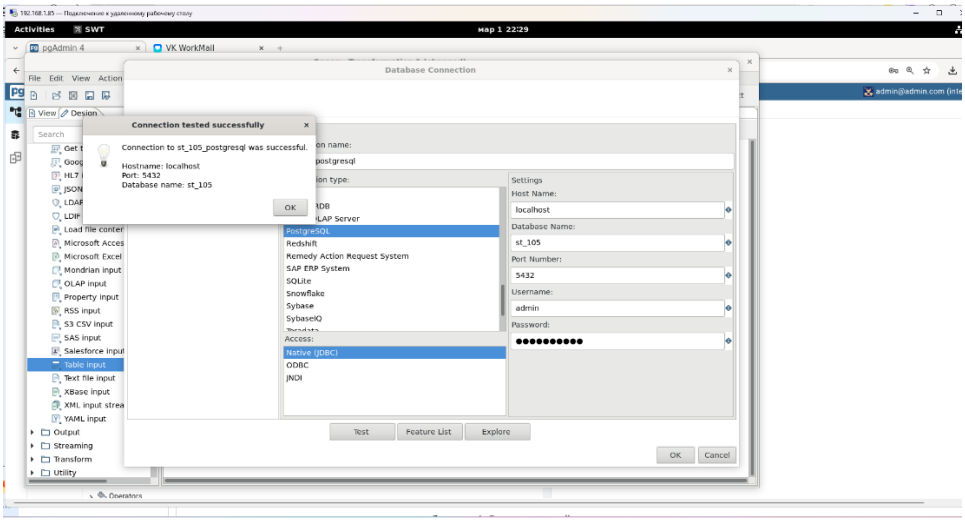

Рисунок 9 – проверка подключения к бд в pgadmin st_105

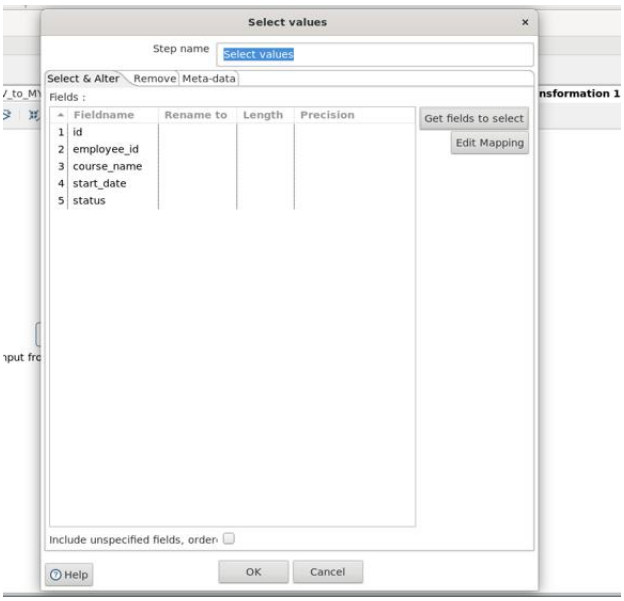

Рисунок 10 –выбор полей из таблицы training postgresql

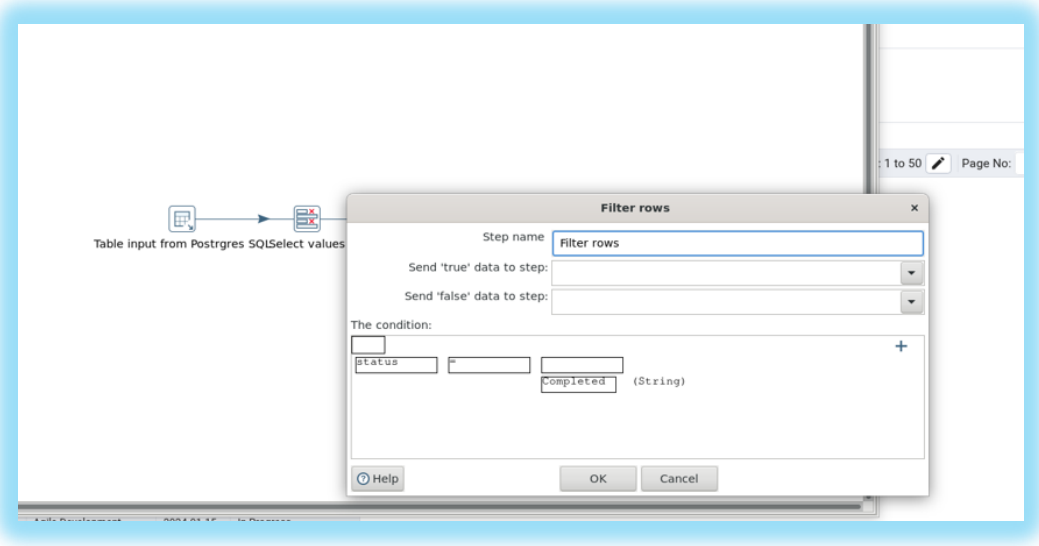

Рисунок 11 – настройка фильтра строк по значению completed

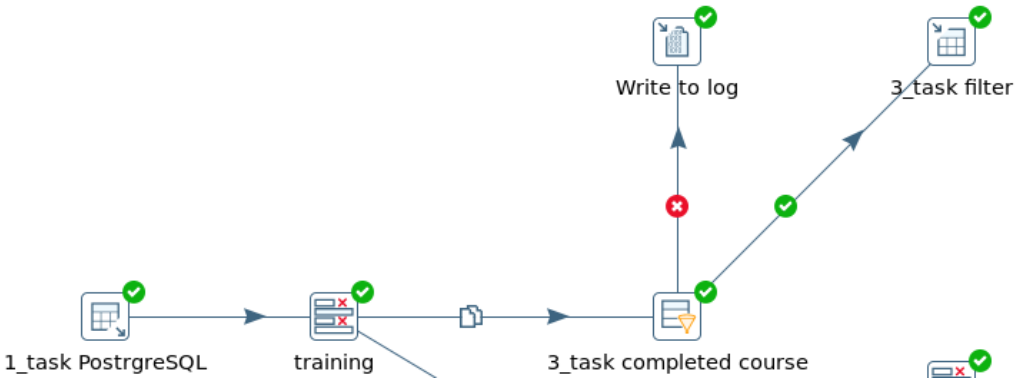

Рисунок 12 – трансформация на этапе задания 3

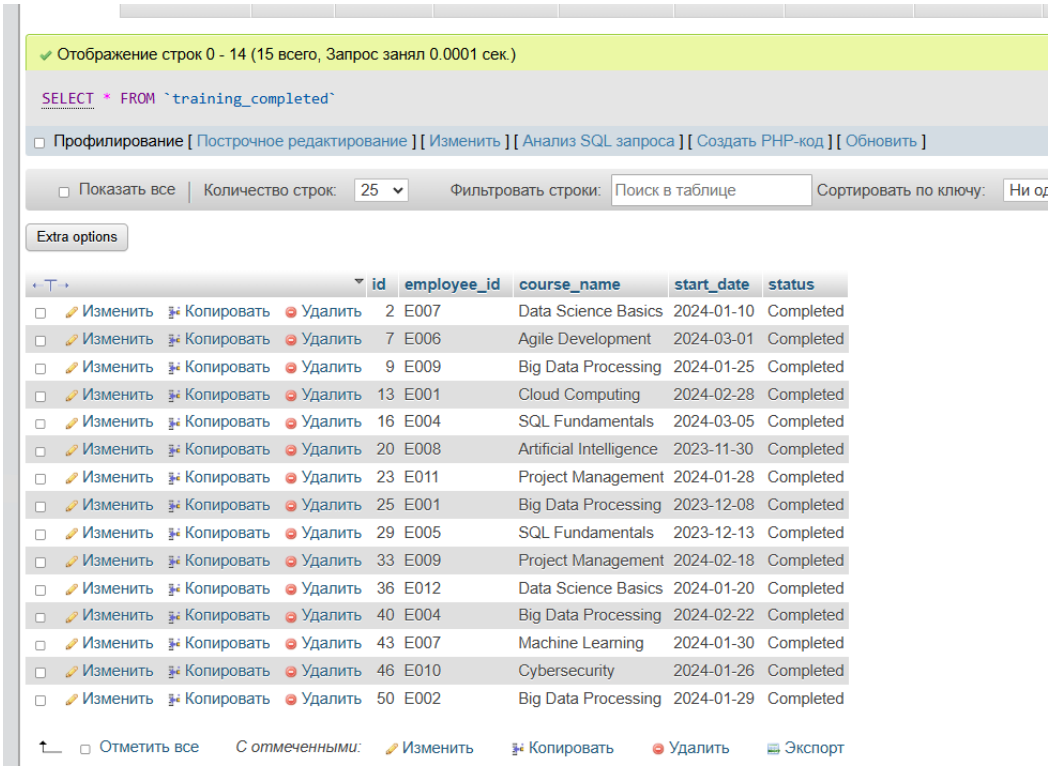

Рисунок 13- итоговый отчёт по 3 заданию

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)

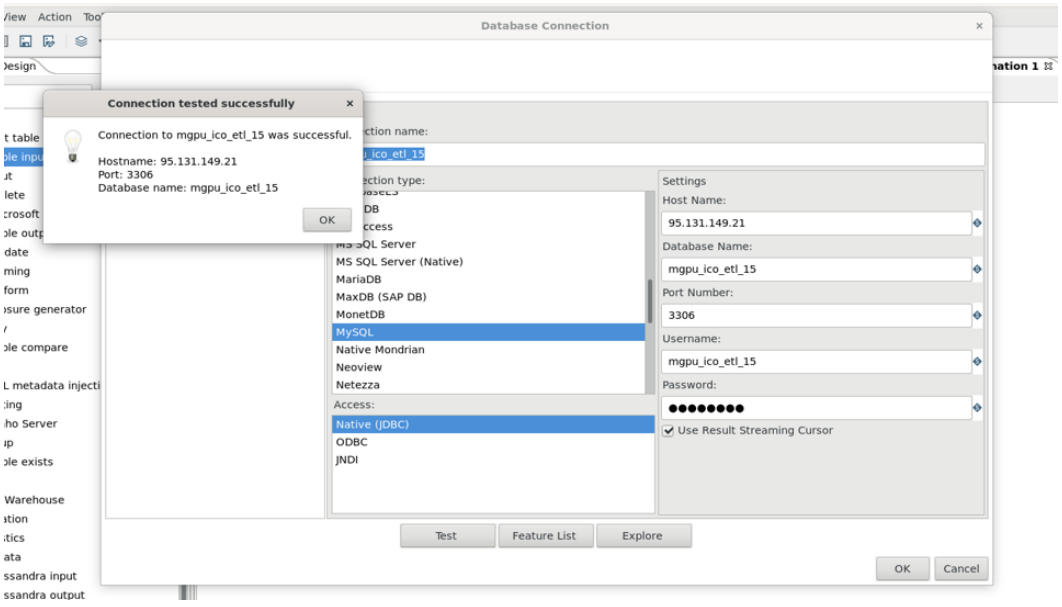

Рисунок 14 – настройка подключения mysql

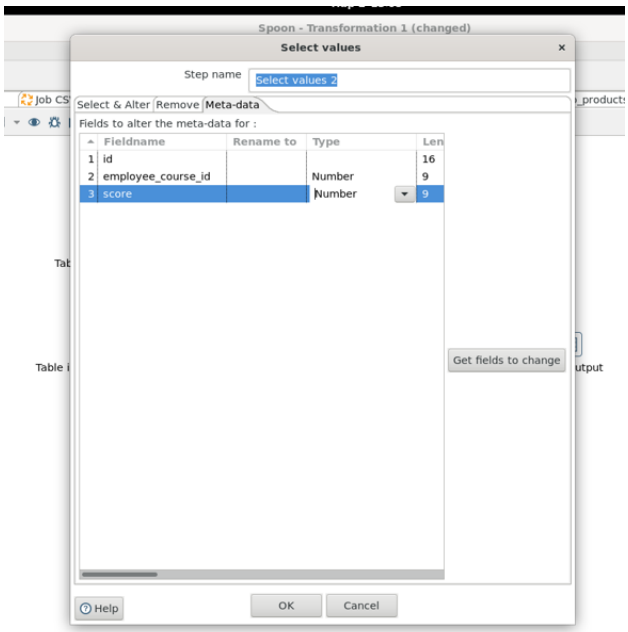

Рисунок 15 – выбор полей из таблицы

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

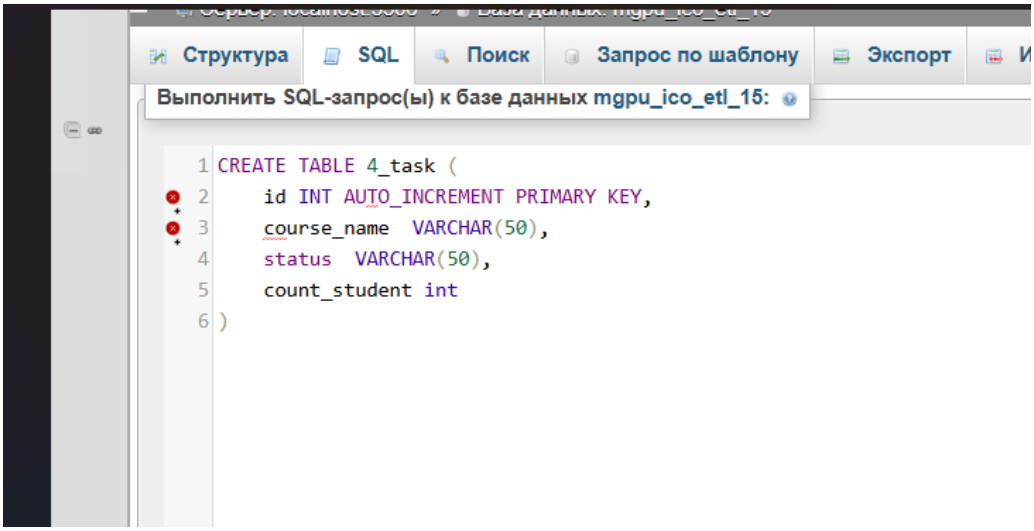

Рисунок 15 – ddl скрипт таблицы 4_task

Настройка узла группировки (рисунок 16)

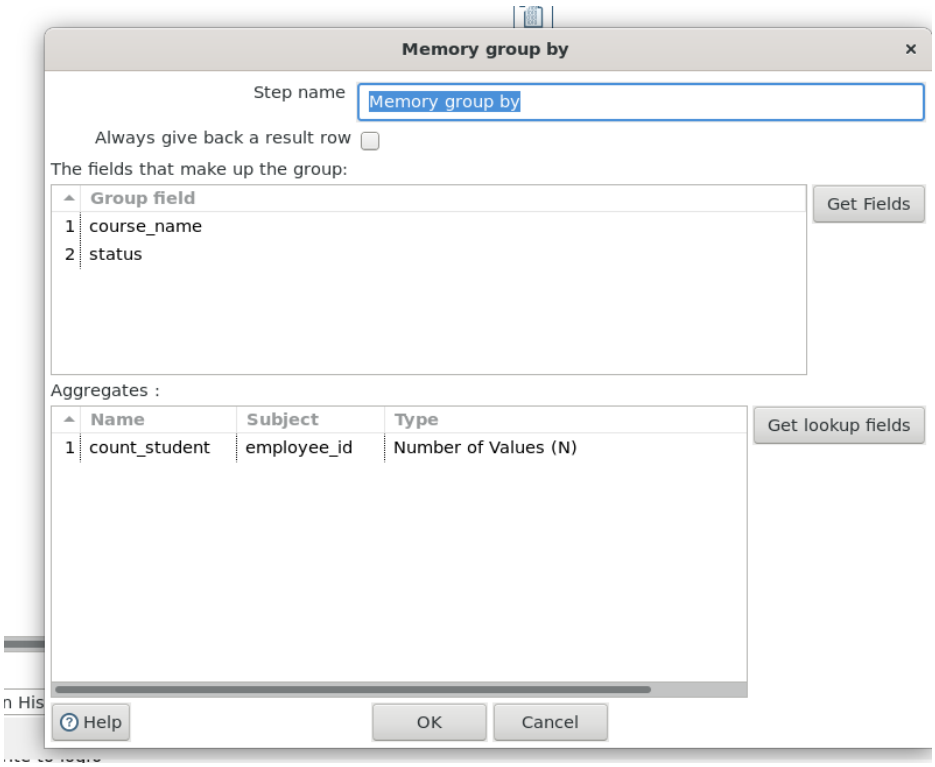

Рисунок 16 – группировка

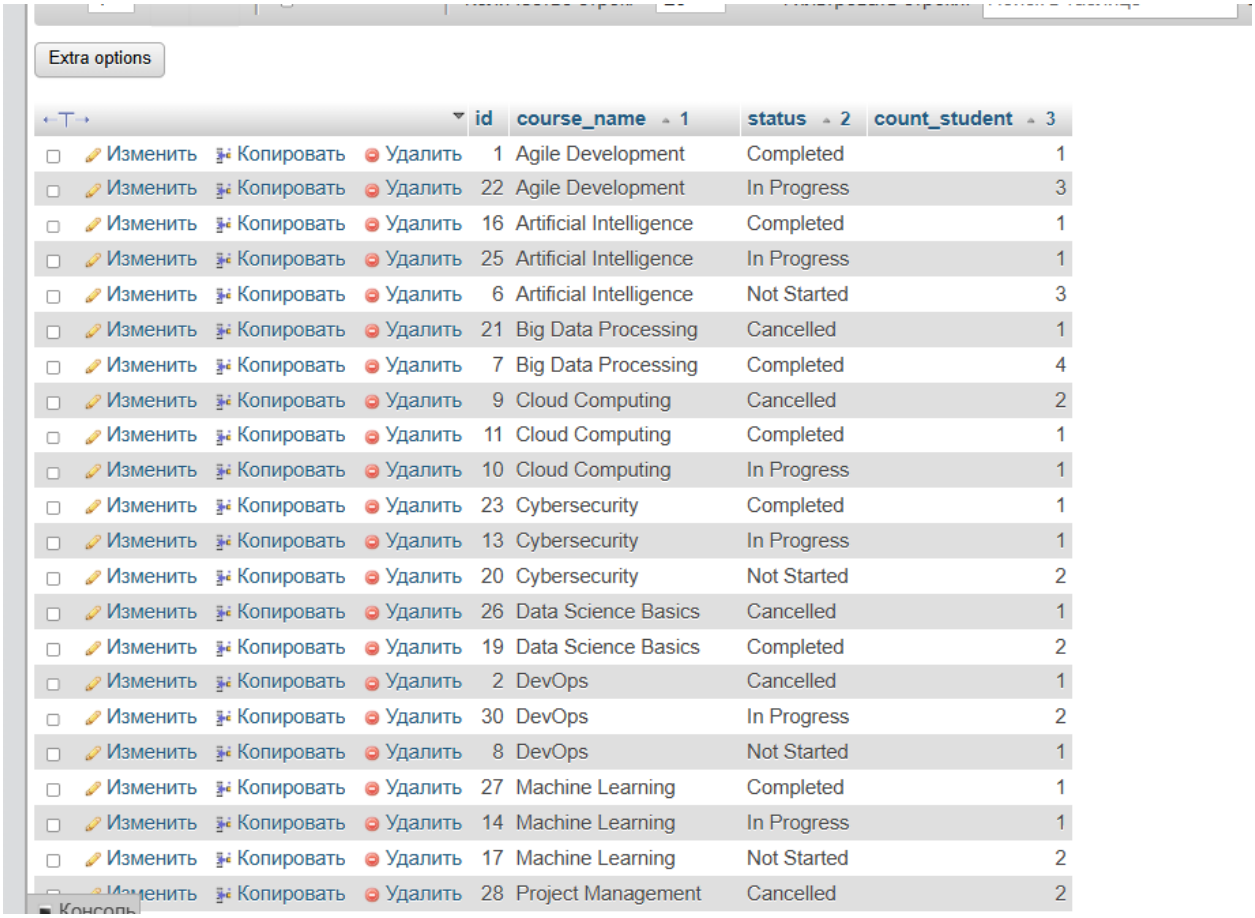

Рисунок 17 – итоговые данные в phpmyadmin

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)

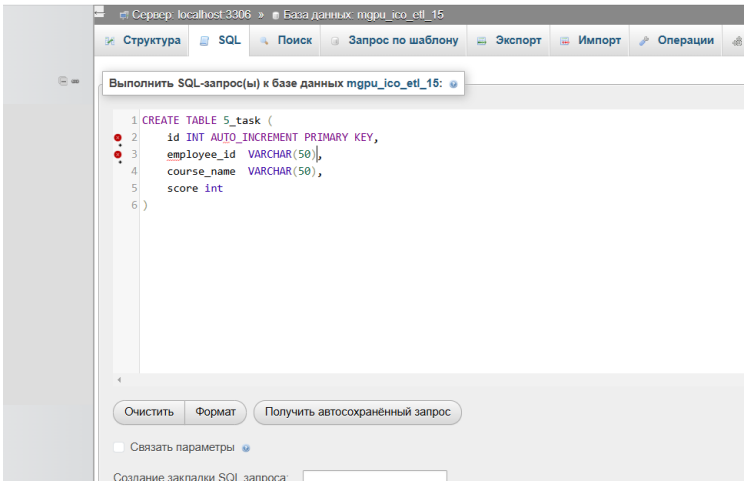

Рисунок 18 – ddl скрипт для 5_task

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

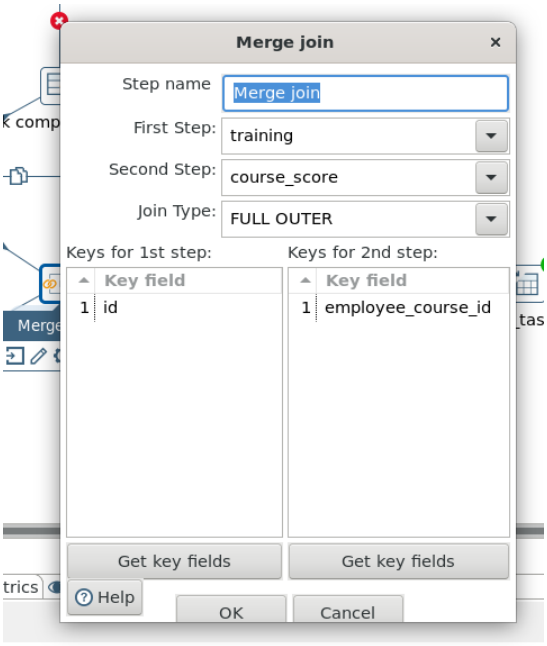

Рисунок 19 – настройка узла объединения

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

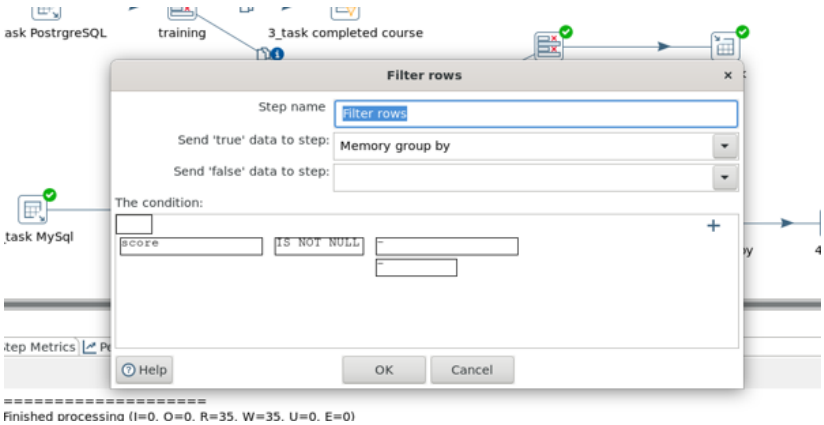

Рисунок 20 – фильтр пустых строк

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

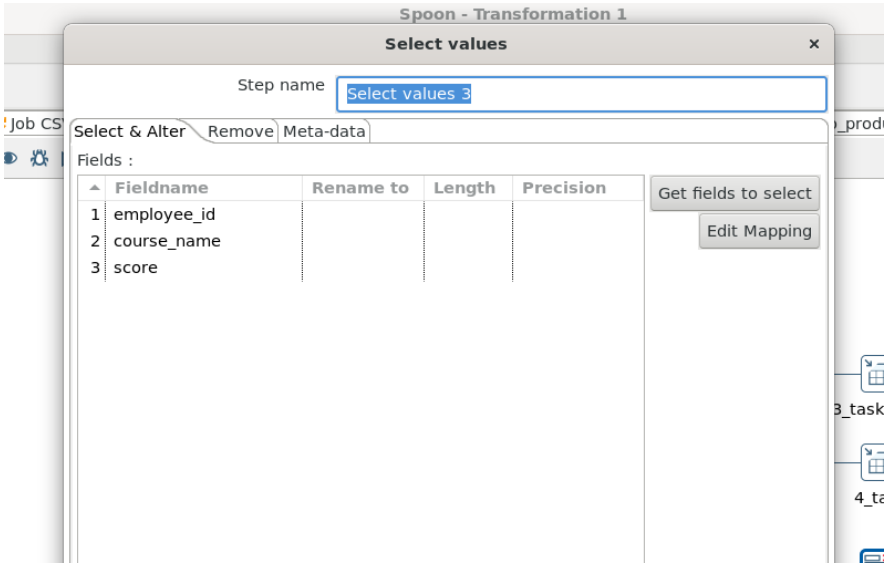

Рисунок 21 – настройка выбора полей

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

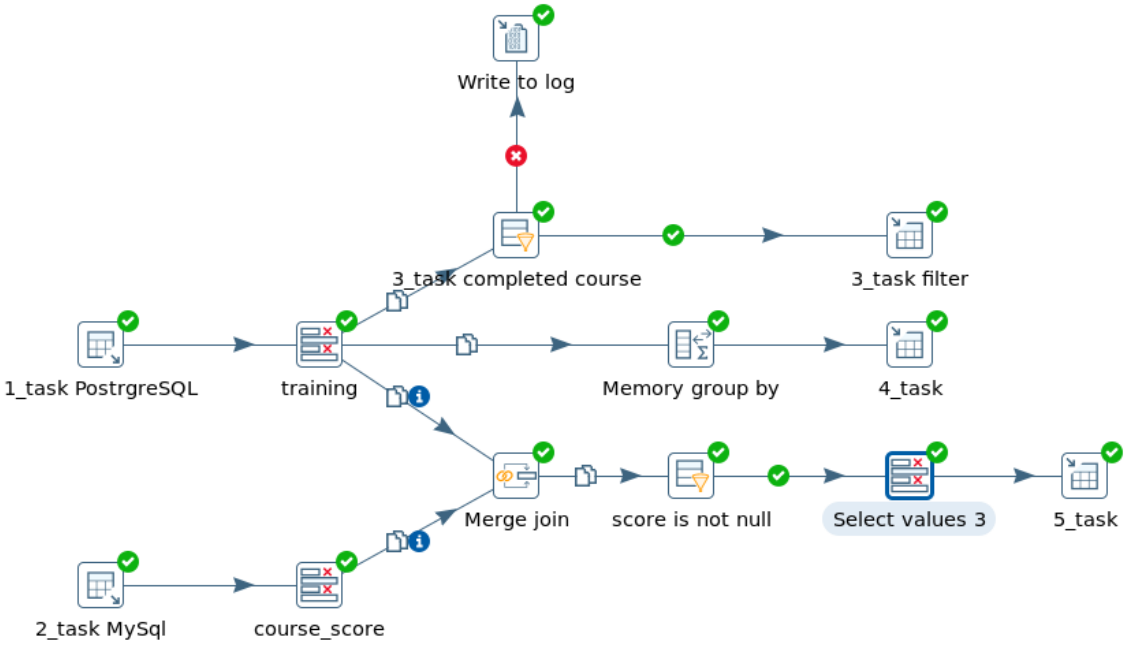

Рисунок 23 – общая трансформация

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