

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

Макарова Екатерина Павловна

ЛАБОРАТОРНАЯ РАБОТА 3

Интеграция данных из разных источников (баз данных)

Проектный практикум по разработке ETL-решений

Направление подготовки

38.03.05 Бизнес-информатика

Профиль подготовки

Аналитика данных и эффективное управление

Курс обучения: 4

Форма обучения: очная

Преподаватель: кандидат технических наук,
доцент Босенко Тимур Муртазович

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.

Задачи для самостоятельной работы:

Задание 1 (PostgreSQL) Создать таблицу projects (id, name, start_date, end_date, budget, status)

Задание 2 (MySQL) Создать таблицу project_tracking с полем completion_percentage

Задание 3 (Pentaho) Фильтр активных проектов

Задание 4 (Pentaho) Расчет освоения бюджета

Задание 5 (Pentaho) Определение процента выполнения

Ход работы:

1. Создание таблицы projects в PostgreSQL

The screenshot shows the PostgreSQL query editor interface. The top bar displays the user 'st_95/admin@MGPU_superset'. Below the toolbar, the 'Query' tab is active, showing the following SQL code:

```
1 CREATE TABLE IF NOT EXISTS projects (  
2     id INTEGER PRIMARY KEY,  
3     name VARCHAR(255) NOT NULL,  
4     start_date DATE NOT NULL,  
5     end_date DATE NOT NULL,  
6     budget DECIMAL(10, 2) NOT NULL,  
7     status VARCHAR(50) NOT NULL  
8 );
```

The 'Messages' tab is also visible, showing the execution result:

```
CREATE TABLE  
  
Query returned successfully in 326 msec.
```

2. Вставка данных (100 записей) в таблицу projects

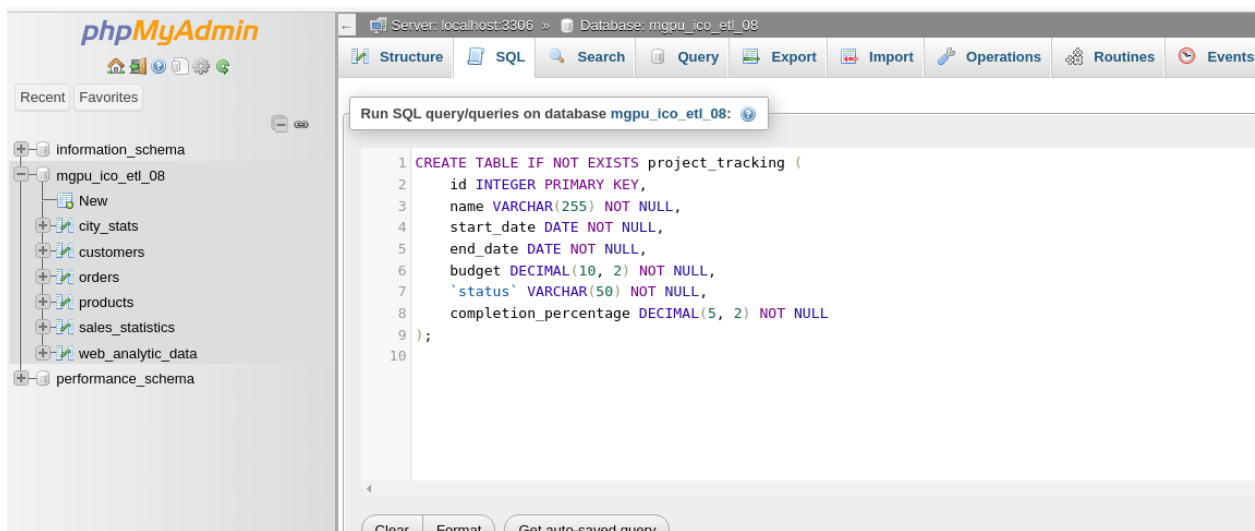
The screenshot shows the PostgreSQL query editor interface. The top bar displays the user 'st_95/admin@MGPU_superset'. Below the toolbar, the 'Query' tab is active, showing the following SQL code:

```
1 INSERT INTO projects (id, name, start_date, end_date, budget, status)  
2 VALUES  
3 (1, 'Project Ocean 123', '2022-01-01', '2022-06-30', 50000.00, 'Active'),  
4 (2, 'Project Sky 456', '2022-03-01', '2022-09-30', 200000.00, 'Completed'),  
5 (3, 'Project Earth 789', '2022-05-01', '2023-02-28', 80000.00, 'On Hold'),  
6 (4, 'Project Fire 101', '2022-07-01', '2023-01-31', 300000.00, 'Active'),  
7 (5, 'Project Water 202', '2022-09-01', '2023-03-31', 400000.00, 'Completed'),  
8 (6, 'Project Air 303', '2022-11-01', '2023-05-31', 600000.00, 'On Hold'),  
9 (7, 'Project Stone 404', '2023-01-01', '2023-07-31', 100000.00, 'Active'),  
10 (8, 'Project Wood 505', '2023-03-01', '2023-09-30', 250000.00, 'Completed'),  
11 (9, 'Project Metal 606', '2023-05-01', '2023-11-30', 350000.00, 'On Hold'),  
12 (10, 'Project Paper 707', '2023-07-01', '2024-01-31', 450000.00, 'Active'),  
13 (11, 'Project Cloud 808', '2023-09-01', '2024-03-31', 550000.00, 'Completed'),  
14 (12, 'Project Sun 909', '2023-11-01', '2024-05-31', 650000.00, 'On Hold'),  
15 (13, 'Project Moon 1010', '2024-01-01', '2024-07-31', 750000.00, 'Active').
```

The 'Messages' tab is also visible, showing the execution result:

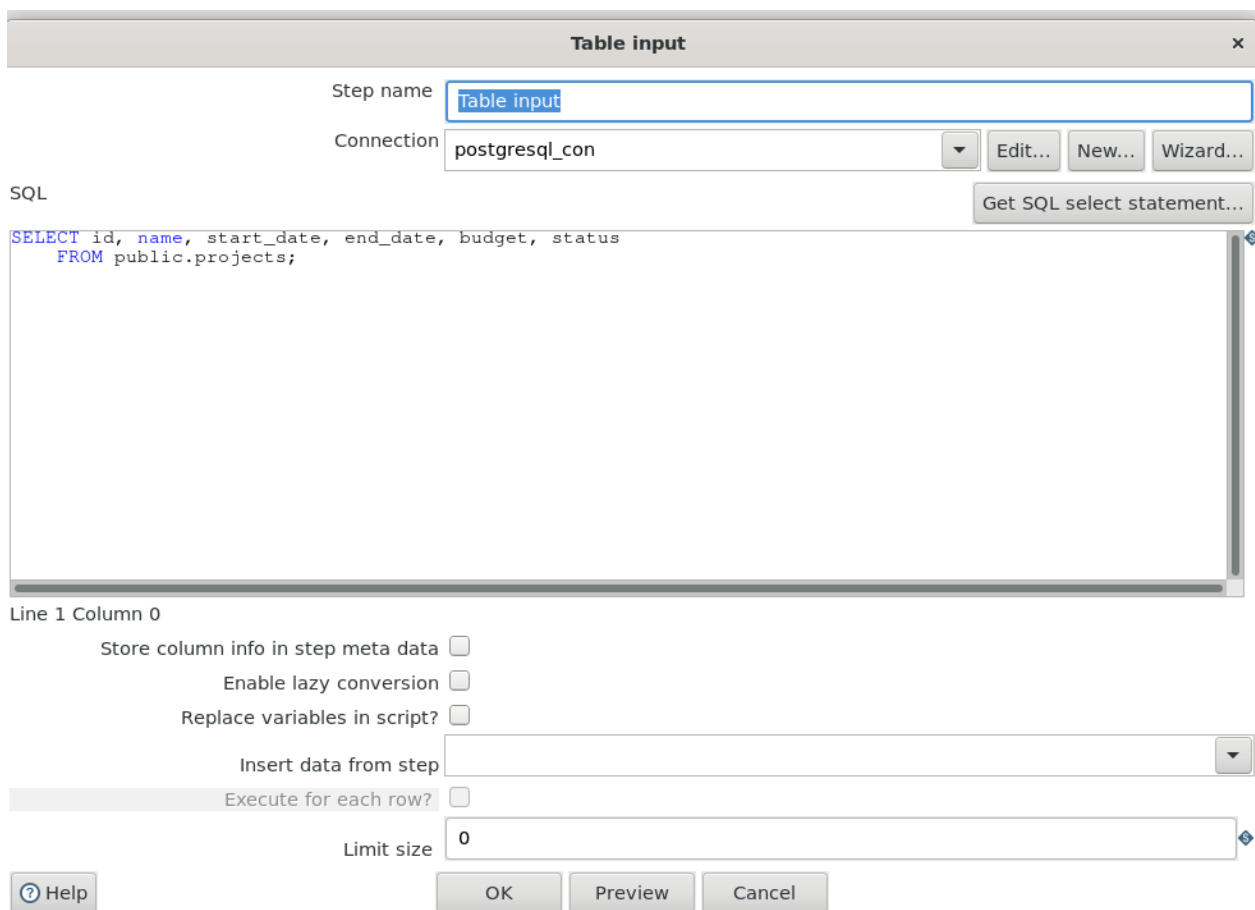
```
INSERT 0 100  
  
Query returned successfully in 321 msec.
```

3. Создание таблицы “project_tracking”



4. Трансформация

1) Создания подключения к PostgreSQL, таблица «project»



2) Фильтр проекта по статусу «Активен»

Filter rows [X]

Step name:

Send 'true' data to step:

Send 'false' data to step:

The condition:

(String)

[?] Help [OK] [Cancel]

3) Для выполнения задания по расчету освоения бюджета необходимо дополнить данными по фактическим затратам:

- Создана новая таблица с данными по фактическим затратам в PostgreSQL

```

1 CREATE TABLE project_actual_costs (
2     cost_id SERIAL PRIMARY KEY,           -- Уникальный ID записи о затратах
3     project_id INT NOT NULL,              -- Ссылка на проект
4     actual_amount DECIMAL(12, 2) NOT NULL CHECK (actual_amount > 0), -- Сумма затрат
5
6     -- Внешний ключ для связи с таблицей projects
7     CONSTRAINT fk_project
8         FOREIGN KEY (project_id)
9         REFERENCES projects(id)
10        ON DELETE CASCADE
11 );
  
```

- Вставляем данные по id и сумме затрат в таблицу

Query	Query History
1	INSERT INTO project_actual_costs (project_id, actual_amount)
2	VALUES
3	(1, 1500.00),
4	(2, 2500.00),
5	(3, 1200.00),
6	(4, 3000.00),
7	(5, 1800.00),
8	(6, 2200.00),
9	(7, 1600.00),
10	(8, 2800.00),
11	(9, 1400.00),
12	(10, 3200.00),
13	(11, 2000.00),
14	(12, 2400.00),
15	(13, 1700.00).

Data Output	Messages	Notifications
INSERT 0 100		

4) Создаём подключение к новой таблице в PostgreSQL

Table input

Step name

psql_actual_zatrat

Connection

postgresql_con

Edit...

New...

Wizard...

SQL

Get SQL select statement...

SELECT cost_id, project_id, actual_amount

FROM public.project_actual_costs;

Line 2 Column 34

Store column info in step meta data

Enable lazy conversion

Replace variables in script?

Insert data from step

Execute for each row?

Limit size

0

Help

OK

Preview

Cancel

5) Объединяем данные по ключу id = project_id

Merge join ×

Step name:

First Step: ▼

Second Step: ▼

Join Type: ▼

Keys for 1st step:

Key field
1 id

Keys for 2nd step:

Key field
1 project_id

Get key fields

Get key fields

? Help

OK Cancel

6) Расчёт процента освоения бюджета: Фактические затраты/Плановые
*100

Calculator ×

Step name:

☒ Throw an error on non existing files

Fields:

	New field	Calculation	Field A	Field B	Field C	Value type	Length	Precision	Remove	Conversion mask	Decimal symbol	Grouping symbol	C
1	budget_development	100 * A / B	actual_amount	budget		None			N				

? Help

OK Cancel

7) Рассмотрим процент завершения проекта через процент освоения бюджета, если процент освоения бюджета = 100, то проект считается завершённым, если меньше, то ставим процент освоения бюджета

Modified JavaScript value

Step name:

JavaScript functions:

- Transform Scripts
- Transform Constants
- Transform Functions
- Input fields
 - id
 - name
 - start_date
 - end_date
 - budget
 - status
 - cost_id

JavaScript:

```
Script 1
//Script here
if (budget_development > 100) {
  completion_percentage = 100;
} else {
  completion_percentage = budget_development;
}
```

Linerr: 0

Compatibility mode? ☐ Optimization level: 9

Fieldname	Rename to	Type	Length	Precision	Replace value 'Fieldname' or 'Rename to'
1 completion_percentage		Number	3		N

Buttons: Help, OK, Cancel, Get variables, Test script

8) Выбираем атрибуты данных, которые будут импортированы в таблицу Mysql

Select values

Step name:

Select & Alter Remove Meta-data

Fields:

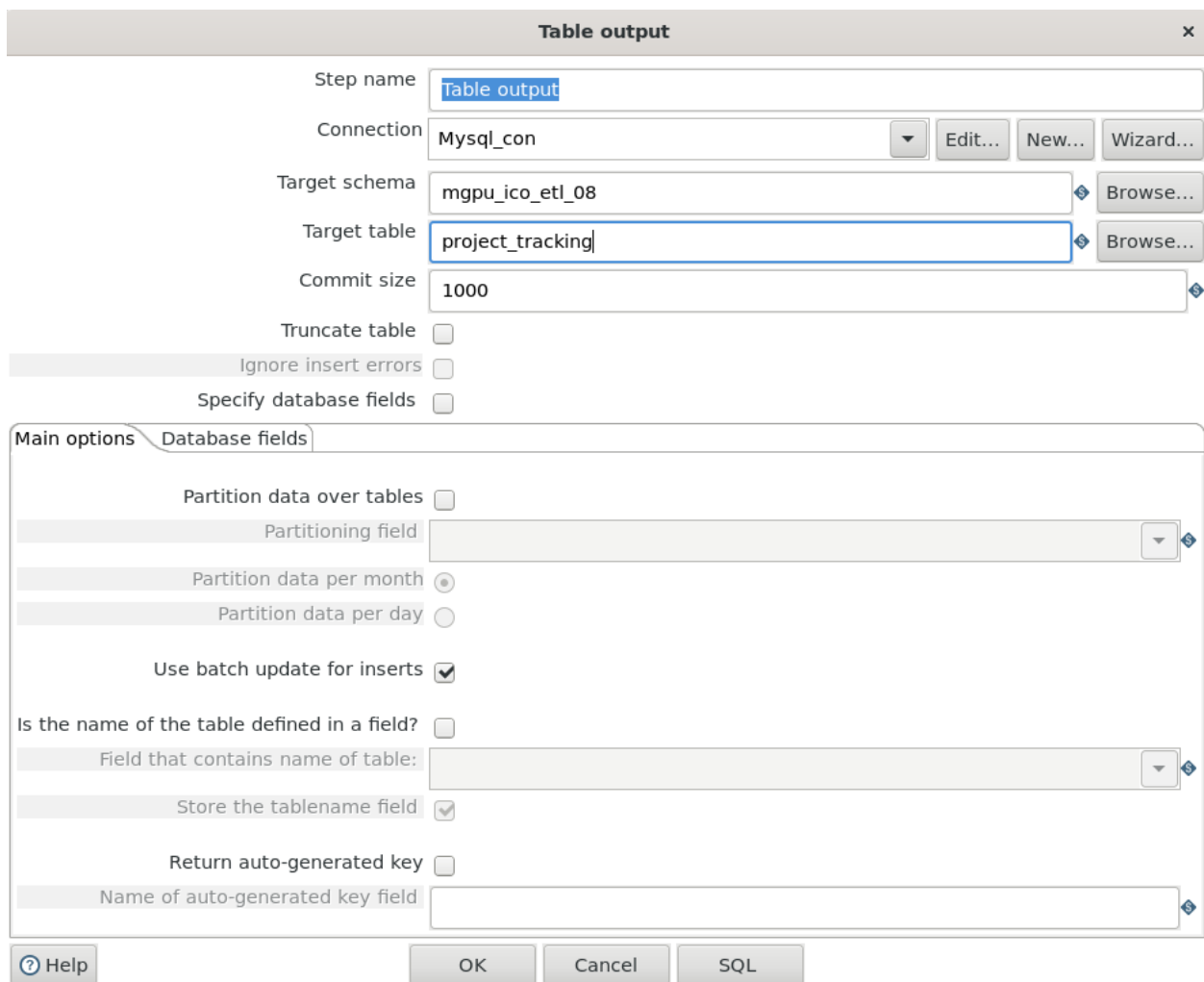
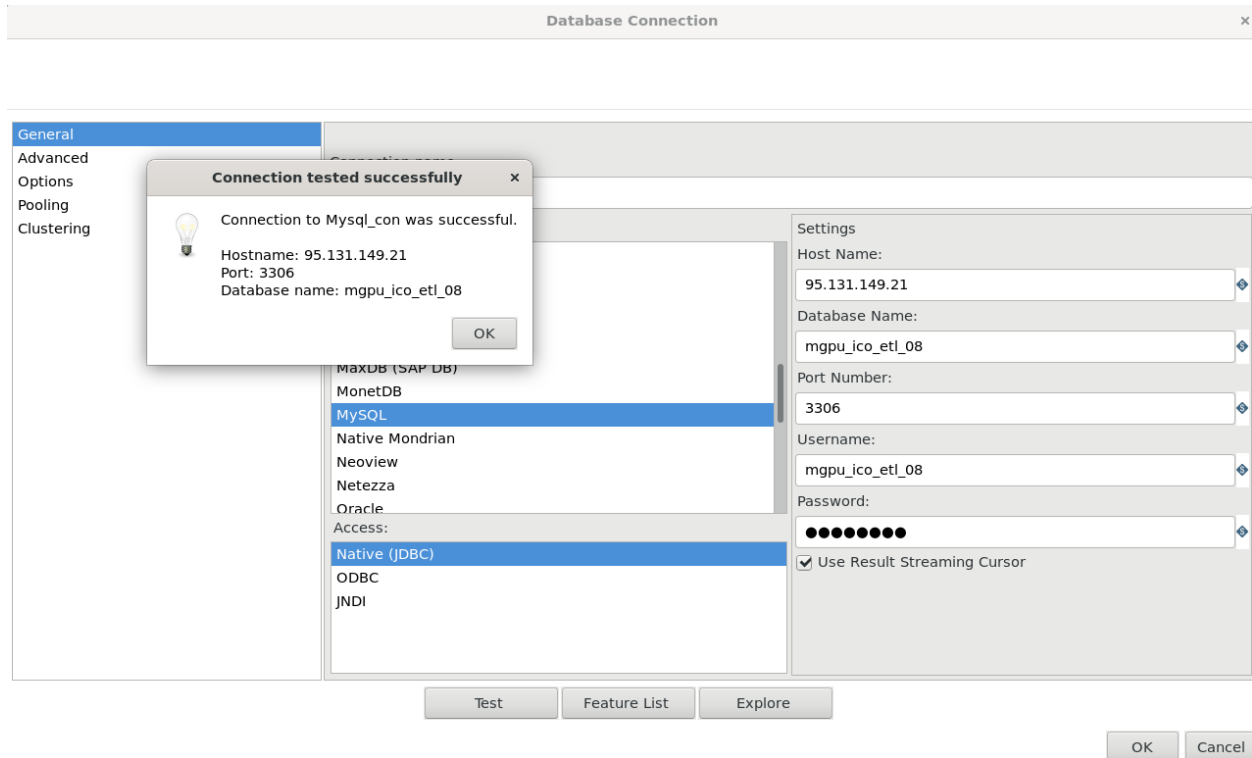
Fieldname	Rename to	Length	Precision
1 id			
2 name			
3 start_date			
4 end_date			
5 budget			
6 status			
7 completion_percentage		2	

Buttons: Get fields to select, Edit Mapping

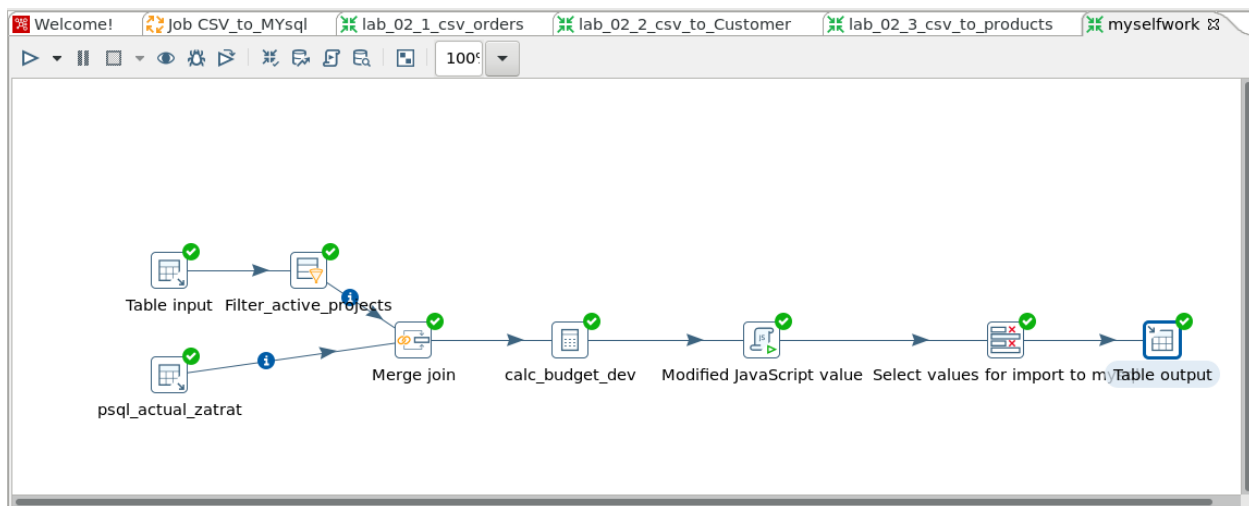
Include unspecified fields, order: ☐

Buttons: Help, OK, Cancel

9) Настраиваем подключение к MySQL



10) Трансформация имеет вид:



11) Настраиваем трансформацию в Job

Transformation

Entry Name:

Transformation:

Options
Logging
Arguments
Parameters

Run configuration:

Execution

☐ Execute every input row
☐ Clear results rows before execution
☐ Clear results files before execution
☒ Wait for remote transformation to complete
☐ Follow local abort to remote transformation
☐ Suppress result data from remote transformation

12) Очищаем таблицы customers, orders, products для импорта данных с помощью Job

☒ MySQL returned an empty result set (i.e. zero rows). (Query took 0.0003 seconds.)

```
TRUNCATE table `customers`;
```

[\[Edit inline \]](#) [\[Edit \]](#) [\[Create PHP code \]](#)

13) Запускаем job, он был успешно выполнен

Job CSV_to_Mysql

Execution Results

Logging History Job metrics Metrics

2025/02/28 15:15:15 - Modified JavaScript value.0 - Optimization level set to 9.
2025/02/28 15:15:15 - Merge join.0 - Finished processing (I=0, O=0, R=134, W=34, U=0, E=0)
2025/02/28 15:15:15 - calc_budget_dev.0 - Finished processing (I=0, O=0, R=34, W=34, U=0, E=0)
2025/02/28 15:15:15 - Modified JavaScript value.0 - Finished processing (I=0, O=0, R=34, W=34, U=0, E=0)
2025/02/28 15:15:15 - Select values for import to mysql.0 - Finished processing (I=0, O=0, R=34, W=34, U=0, E=0)
2025/02/28 15:15:17 - Table output.0 - Finished processing (I=0, O=34, R=34, W=34, U=0, E=0)

14) Проверяем импортированные данные из PostgreSQL

Server: localhost:3306 Database: mgpu_ico_etl_08 Table: project_tracking

Showing rows 0 - 24 (34 total, Query took 0.0007 seconds.)

```
SELECT * FROM `project_tracking`
```

Profiling [Edit inline] [Edit] [Explain SQL] [Create PHP code] [Refresh]

1 > >> Show all Number of rows: 25 Filter rows: Search this table Sort by key: None

				id	name	start_date	end_date	budget	status	completion_percentage
<input type="checkbox"/>	Edit	Copy	Delete	1	Project Ocean 123	2022-01-01	2022-06-30	50000.00	Active	3.00
<input type="checkbox"/>	Edit	Copy	Delete	4	Project Fire 101	2022-07-01	2023-01-31	300000.00	Active	1.00
<input type="checkbox"/>	Edit	Copy	Delete	7	Project Stone 404	2023-01-01	2023-07-31	100000.00	Active	1.60
<input type="checkbox"/>	Edit	Copy	Delete	10	Project Paper 707	2023-07-01	2024-01-31	450000.00	Active	0.71
<input type="checkbox"/>	Edit	Copy	Delete	13	Project Moon 1010	2024-01-01	2024-07-31	750000.00	Active	0.23
<input type="checkbox"/>	Edit	Copy	Delete	16	Project Planet 1313	2024-07-01	2025-01-31	1000000.00	Active	0.31
<input type="checkbox"/>	Edit	Copy	Delete	19	Project Space 1616	2025-01-01	2025-07-31	1150000.00	Active	0.14
<input type="checkbox"/>	Edit	Copy	Delete	22	Project Energy 1919	2025-07-01	2026-01-31	1300000.00	Active	0.24
<input type="checkbox"/>	Edit	Copy	Delete	25	Project Motion 2222	2026-01-01	2026-07-31	1450000.00	Active	0.12
<input type="checkbox"/>	Edit	Copy	Delete	28	Project Sound 2525	2026-07-01	2027-01-31	1600000.00	Active	0.20
<input type="checkbox"/>	Edit	Copy	Delete	31	Project Field 2828	2027-01-01	2027-07-31	1750000.00	Active	0.09
<input type="checkbox"/>	Edit	Copy	Delete	34	Project Evolution 3131	2027-07-01	2028-01-31	1900000.00	Active	0.17
<input type="checkbox"/>	Edit	Copy	Delete	37	Project Future 3434	2028-01-01	2028-07-31	2050000.00	Active	0.08
<input type="checkbox"/>	Edit	Copy	Delete	40	Project Development 3737	2028-07-01	2029-01-31	2200000.00	Active	0.14
<input type="checkbox"/>	Edit	Copy	Delete	43	Project Science 4040	2029-01-01	2029-07-31	2350000.00	Active	0.07
<input type="checkbox"/>	Edit	Copy	Delete	46	Project Exploration 4343	2029-07-01	2030-01-31	2500000.00	Active	0.13
<input type="checkbox"/>	Edit	Copy	Delete	49	Project Journey 4646	2030-01-01	2030-07-31	2650000.00	Active	0.06
<input type="checkbox"/>	Edit	Copy	Delete	52	Project Destination 4949	2030-07-01	2031-01-31	2800000.00	Active	0.12

Выводы по работе:

В ходе работы были выполнены поставленные задачи