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

**Лабораторная работа №3**

**Интеграция данных из нескольких источников. Обработка и согласование данных из разных источников**

Выполнил(а): st\_98 группа: АДЭУ-211

Преподаватель:

Москва

2025

**Цель работы:** получить практические навыки интеграции, обработки и согласования данных из различных источников с использованием Python и его библиотек.

**Задачи:**

* Изучить методы чтения данных из разных источников.
* Освоить техники обработки и очистки данных.
* Научиться согласовывать данные из разных источников.
* Реализовать сохранение обработанных данных.
* Необходимо нарисовать верхнеуровневую архитектуру аналитического решения в <https://draw.io/>

**Вариант 11.**

Интеграция данных о закупках:

* MySQL база данных поставщиков.
* Excel файл с ценовыми предложениями.
* CSV файл с историей закупок.

Задача: оптимизировать закупочные процессы.

**Ход работы:**

1. Первым этапом необходимо спроектировать верхнеуровневую архитектуру аналитического решения. Данная архитектура представлена на рисунке 1.

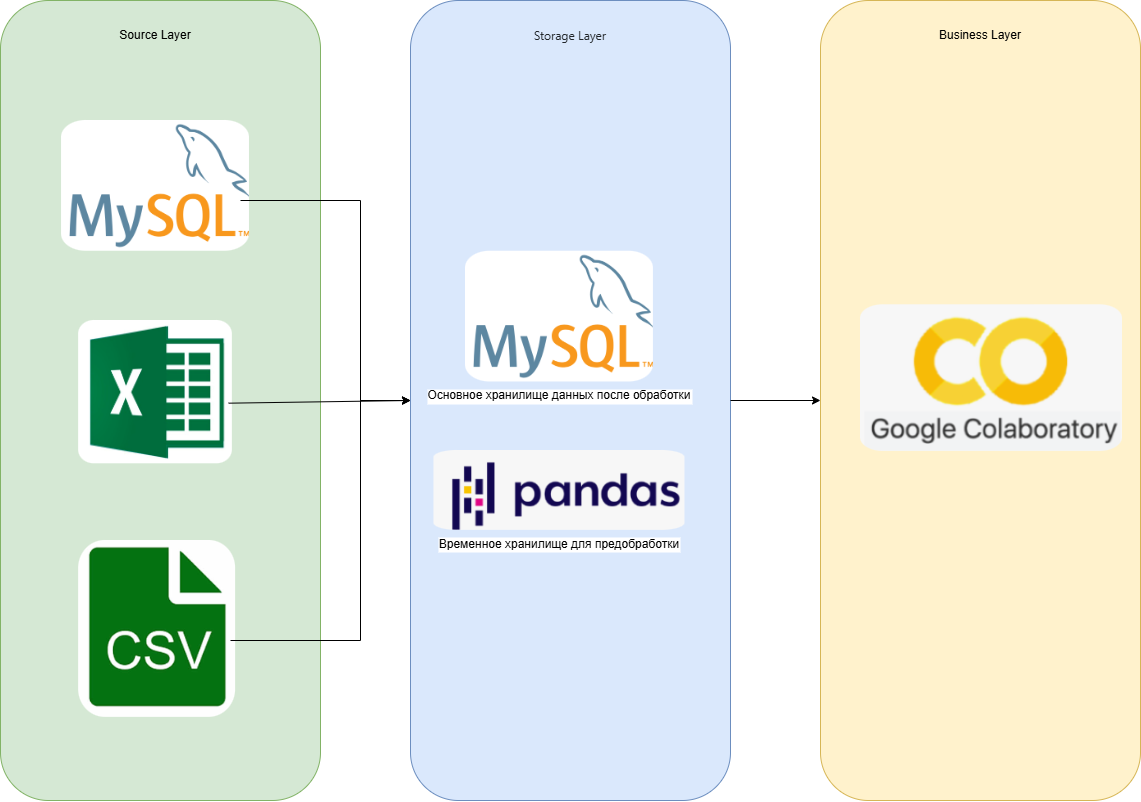


Рисунок 1 – Верхнеуровневая архитектура

1. Далее нужно создать таблицу с поставщиками в MySQL (рисунок 2).

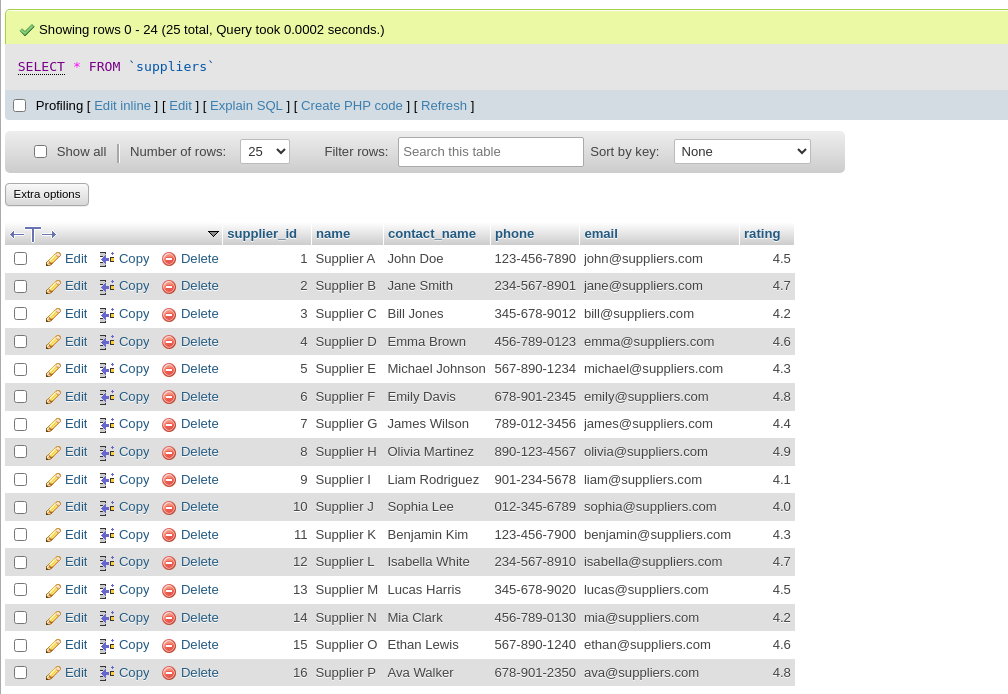


Рисунок 2 – Таблица с поставщиками

1. Следующим этапом в Google Colab надо загрузить необходимые библиотеки, а именно: pandas, numpy, sqlalchemy, mysql.connection, psycopg2, настроить подключение к mysql к таблице с поставщиками (рисунок 3).

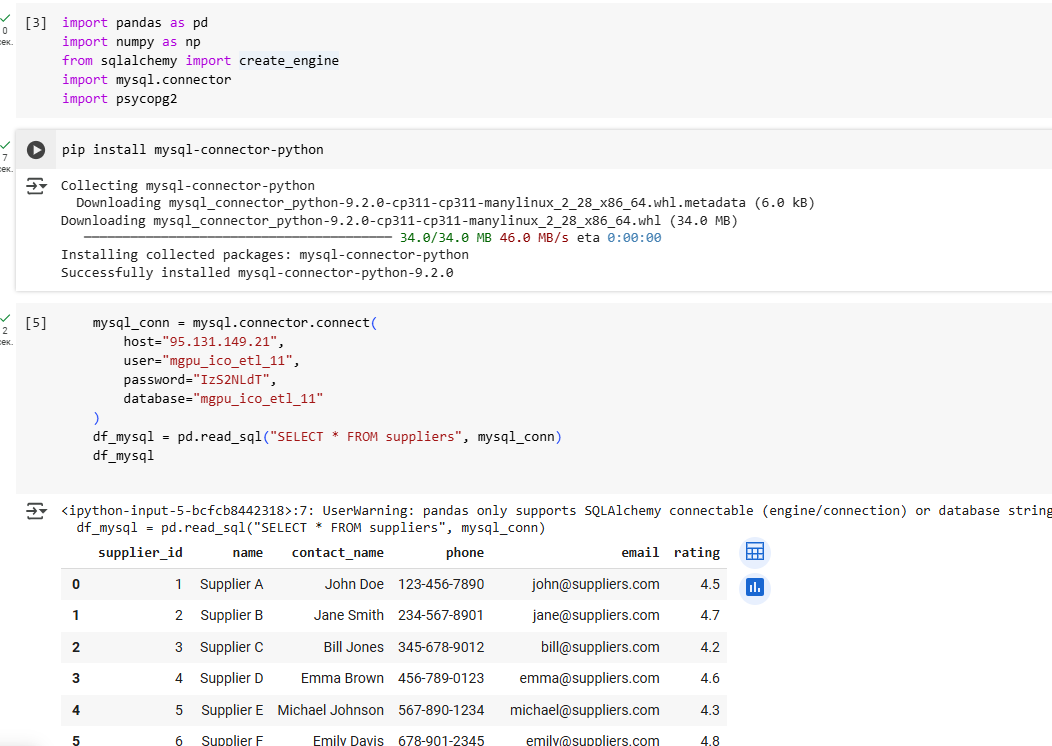


Рисунок 3 – Импорт библиотек и загрузка таблицы из MySQL

1. Далее нужно удалить дубликаты и посмотреть типы данных. Как можно заметить на рисунке 4, они соответствуют действительным. Те же самые действия делаются и с csv-файлом.

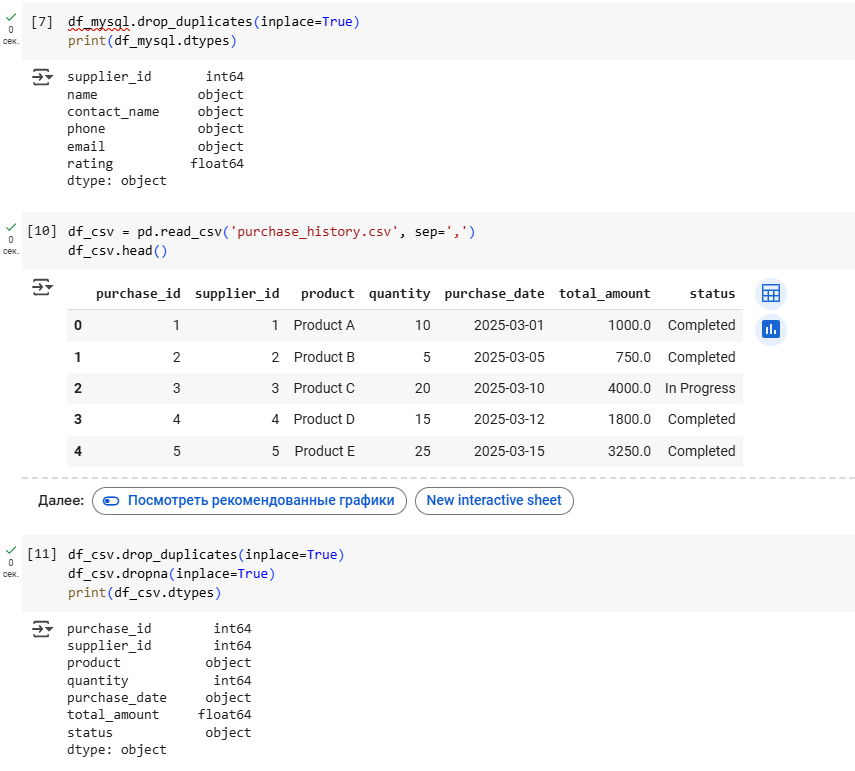


Рисунок 4 – Удаление дубликатов, повторяющихся значений, загрузка csv-файла

1. Для следующего этапа был создан excel-файл, его необходимо загрузить и удалить дубликаты, повторяющиеся значения (рисунок 5).

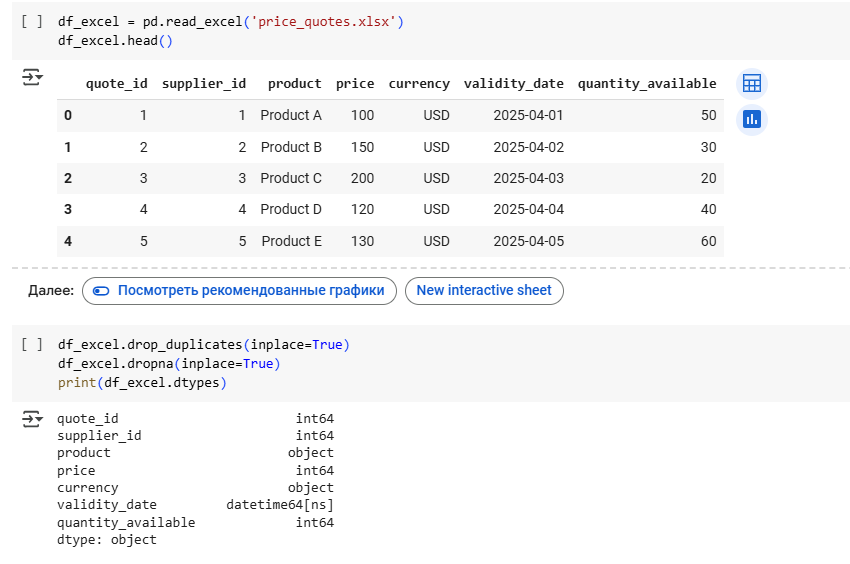


Рисунок 5 – Загрузка excel-файла и предобработка

1. Далее нужно объединить три предобработанные таблицы между собой путем внутреннего соединения (рисунок 6).

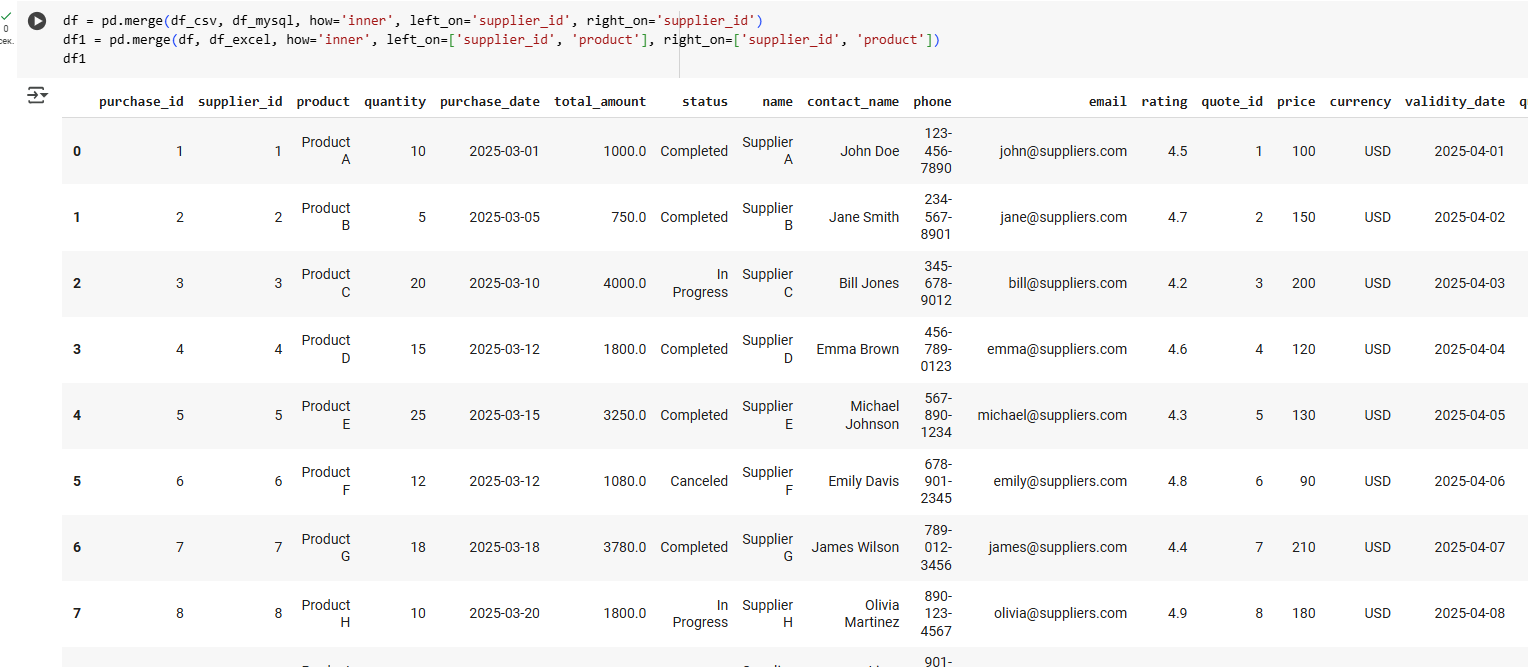


Рисунок 6 – Объединение трех таблиц между собой

1. Затем нужно оптимизировать закупочные процессы. Для этого нужно определить эффективность поставщика по количеству заверенных заказов, сравнить цены с историческими закупками и выявить наилучшие предложения по каждому продукту (рисунок 7).

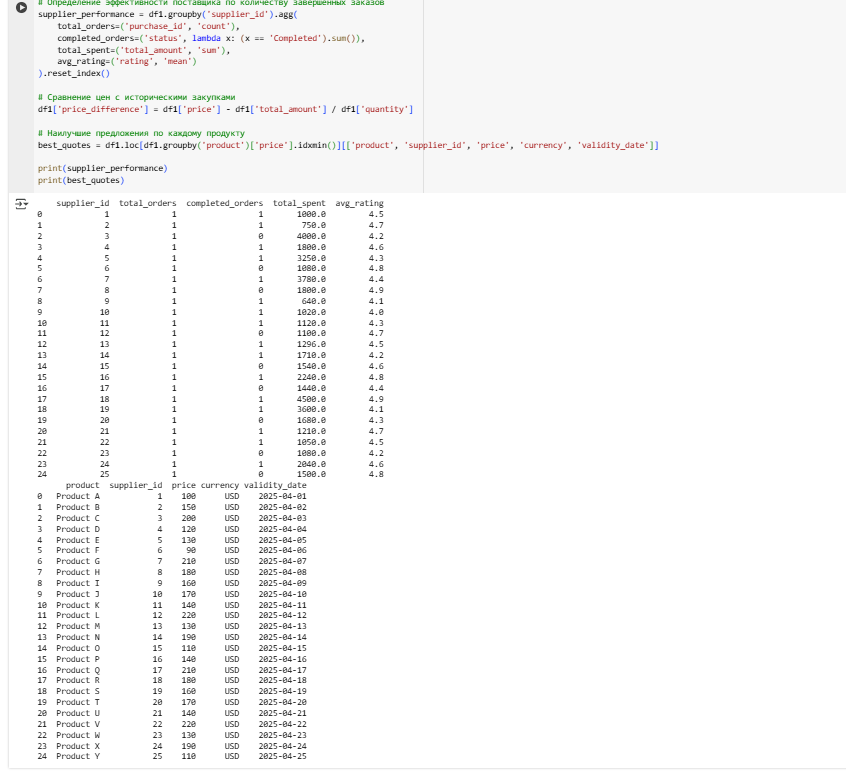


Рисунок 7 – Оптимизация закупочных процессов

1. Далее построим график распределения разницы в ценах. В 64% случаев разницы нет (рисунок 8).

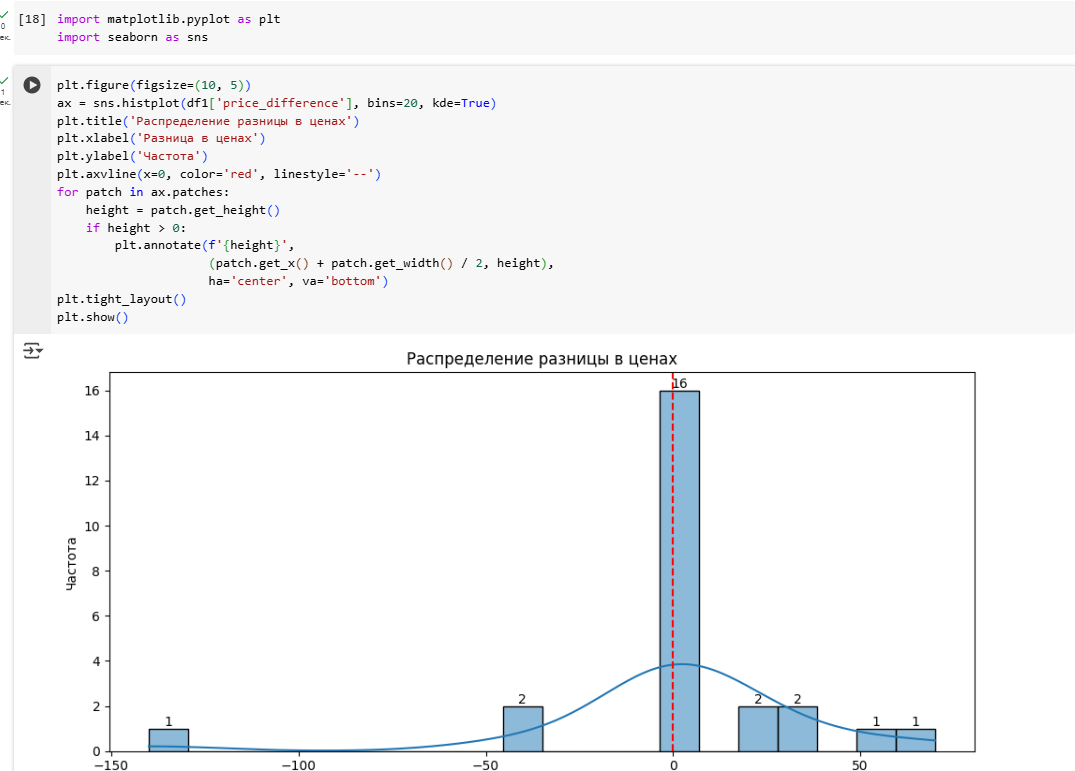


Рисунок 8 – График распределения разницы в ценах

1. Выгружаем полученную таблицу в MySQL (рисунок 9).



Рисунок 9. Выгрузка в MySQL

1. Как можно увидеть на рисунке 10, выгрузка в MySQL прошла успешно.

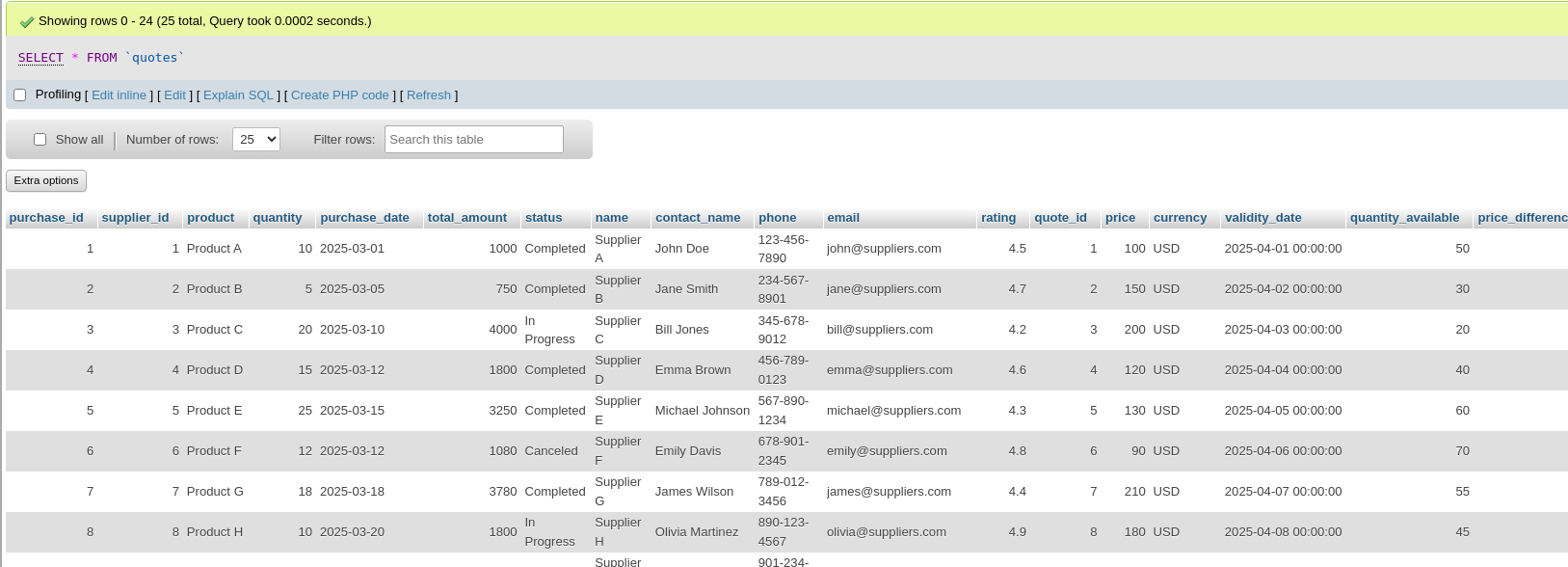


Рисунок 10. Выгрузка в MySQL

**Выводы:**

1. Была спроектирована верхнеуровневая архитектура аналитического решения.

2. Изучены методы чтения данных из разных источников.

3. Освоены техники очистки и обработки данных.

4. Реализовано сохранение обработанных данных.