Отчет по проекту "Приложение для автоматизации ETL процессов"

1. Спецификация требований (SRS) по стандарту ISO/IEC/IEEE 29148

1.1 Введение

1.1.1 Назначение

Программа "Приложение для автоматизации ETL процессов" предназначена для управления ETL (Extract, Transform, Load) процессами через графический интерфейс, позволяя пользователям:

- Подключаться к различным источникам данных
- Очищать и преобразовывать данные
- Создавать структуры хранения данных
- Настраивать процессы переноса данных
- Выполнять запросы к данным

1.1.2 Область применения

Приложение может использоваться для:

- Миграции данных между системами
- Создания хранилищ данных
- Автоматизации процессов обработки данных
- Обучения принципам ETL

1.2 Описание продукта

1.2.1 Функциональность

Основные функции:

1. Управление подключениями к БД (PostgreSQL, MySQL)
2. Загрузка данных из CSV и таблиц БД
3. Очистка данных (удаление дубликатов)
4. Создание и модификация таблиц
5. Настройка маппинга полей

6. Реализация SCD2 (медленно меняющихся измерений)
7. Выполнение SQL-запросов

1.2.2 Пользователи

- Аналитики данных
- Разработчики ETL-процессов
- Администраторы БД
- Студенты, изучающие ETL

1.3 Требования

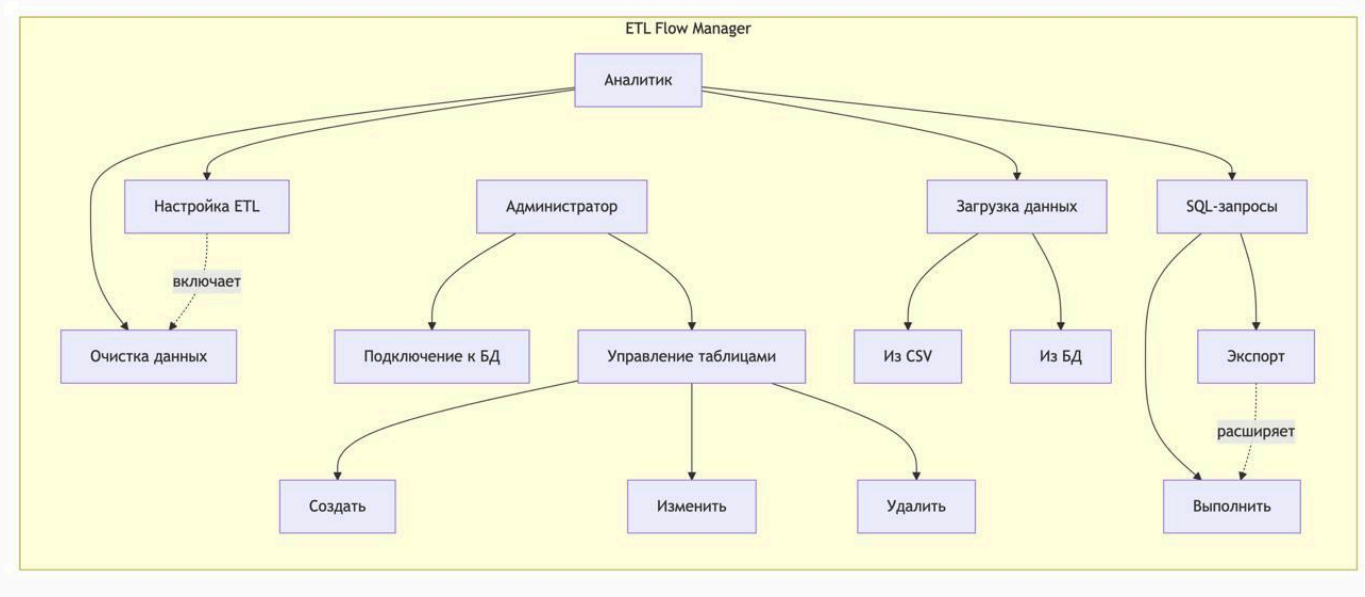
1.3.1 Функциональные требования

1. Подключение к БД:
    - Поддержка PostgreSQL и MySQL
    - Проверка соединения
2. Загрузка данных:
    - Из CSV файлов
    - Из таблиц БД
3. Очистка данных:
    - Удаление полных дубликатов
4. Управление таблицами:
    - Создание новых таблиц
    - Добавление колонок
5. Настройка ETL:
    - Маппинг полей источника и назначения
    - Настройка SCD2
6. Выполнение запросов:
    - Произвольные SQL-запросы
    - Просмотр результатов

1.3.2 Нефункциональные требования

1. Интерфейс: графический (PyQt5)
2. Язык: Python 3.8+
3. Зависимости: pandas, SQLAlchemy, PyQt5
4. Кроссплатформенность: Windows, Linux, macOS

1.4 Диаграмма прецедентов



2. Этап проектирования

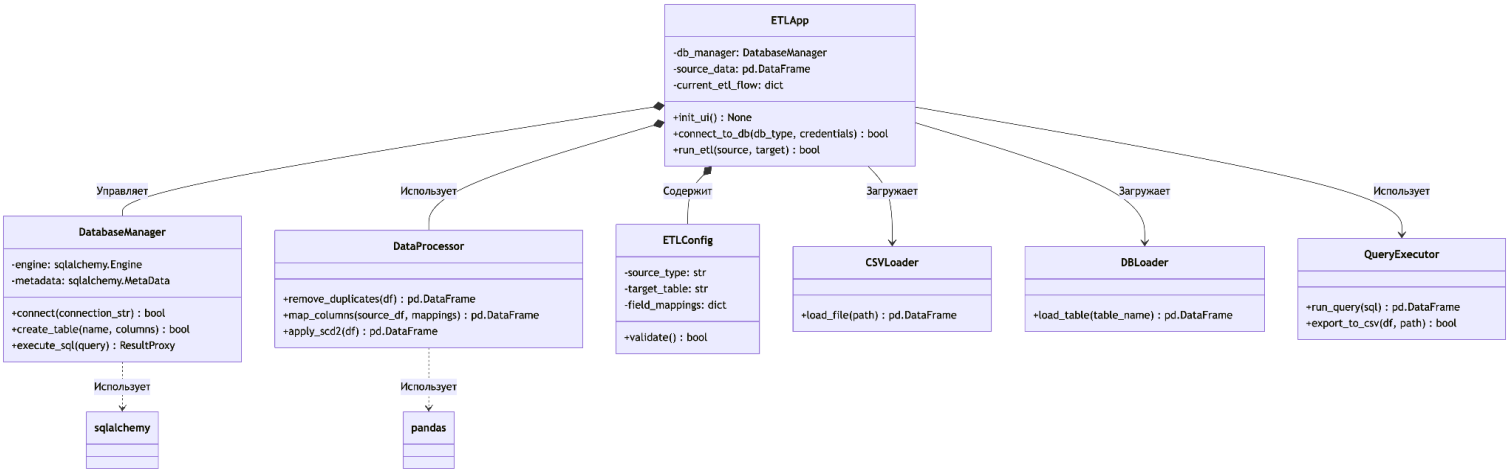
2.1 Таблица прецедентов

ID	Название прецедента	Актор	Описание	Предусловия	Постусловия
1	Подключиться к БД	Администратор БД	Установка соединения с PostgreSQL/MySQL	Параметры подключения корректны	Статус "Connected" в интерфейсе
2	Загрузить данные из CSV	Аналитик	Импорт данных из CSV-файла в систему	Файл существует и корректен	Данные отображаются в таблице предпросмотра

3	Загрузить данные из таблицы БД	Аналитик	Извлечение данных из выбранной таблицы БД	Подключение к БД активно	Данные отображаются в интерфейсе
4	Удалить дубликаты	Аналитик	Очистка данных от полных дубликатов строк	Данные загружены в систему	Число строк уменьшено
5	Создать таблицу	Администратор БД	Создание новой таблицы с указанными колонками и ограничениями	Подключение к БД активно	Таблица отображается в списке существующих
6	Добавить колонку	Администратор БД	Добавление новой колонки в существующую таблицу	Таблица существует	Структура таблицы обновлена
7	Настроить ETL-процесс	Аналитик	Определение маппинга полей и параметров загрузки (включая SCD2)	Источник и целевая таблица выбраны	Конфигурация сохранена
8	Выполнить SQL-запрос	Аналитик/Администратор	Выполнение произвольного SQL-запроса к БД	Подключение к БД активно	Результаты отображаются в таблице

9	Экспорт результатов в CSV	Аналитик	Сохранение результатов запроса в CSV-файл

2.2 Диаграмма концептуальных классов



2.3 Диаграмма последовательностей

Диаграмма последовательности "Создание таблицы"

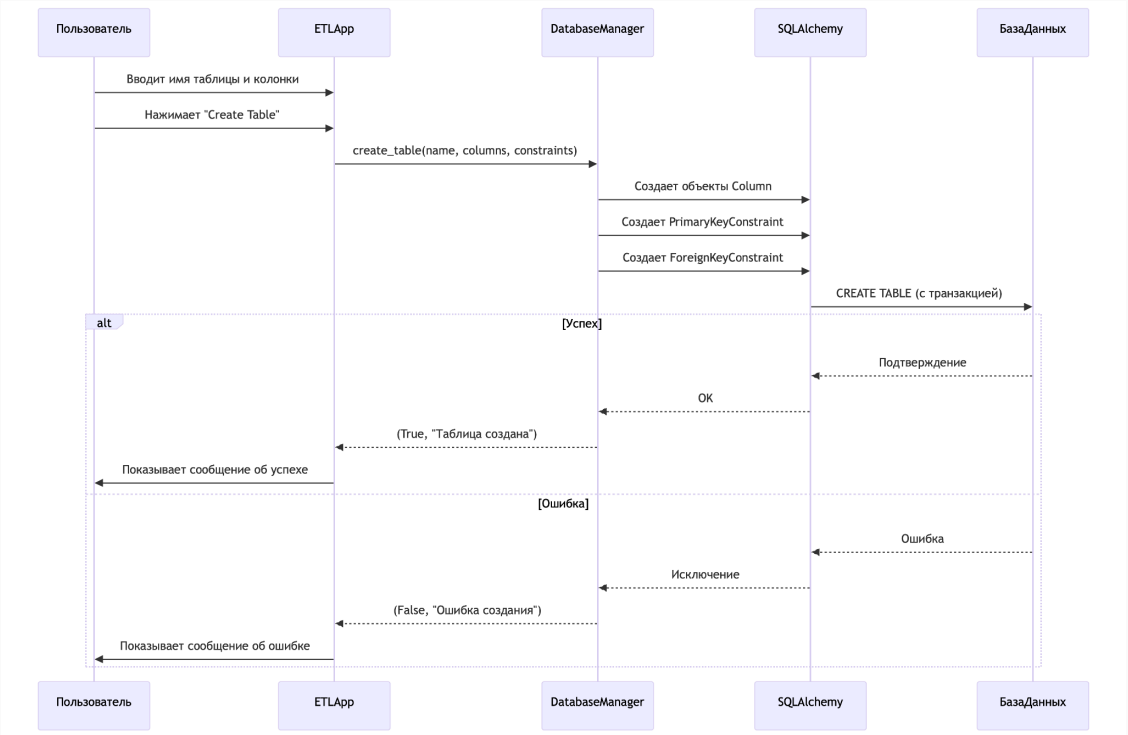


Диаграмма последовательности "Добавление колонки"

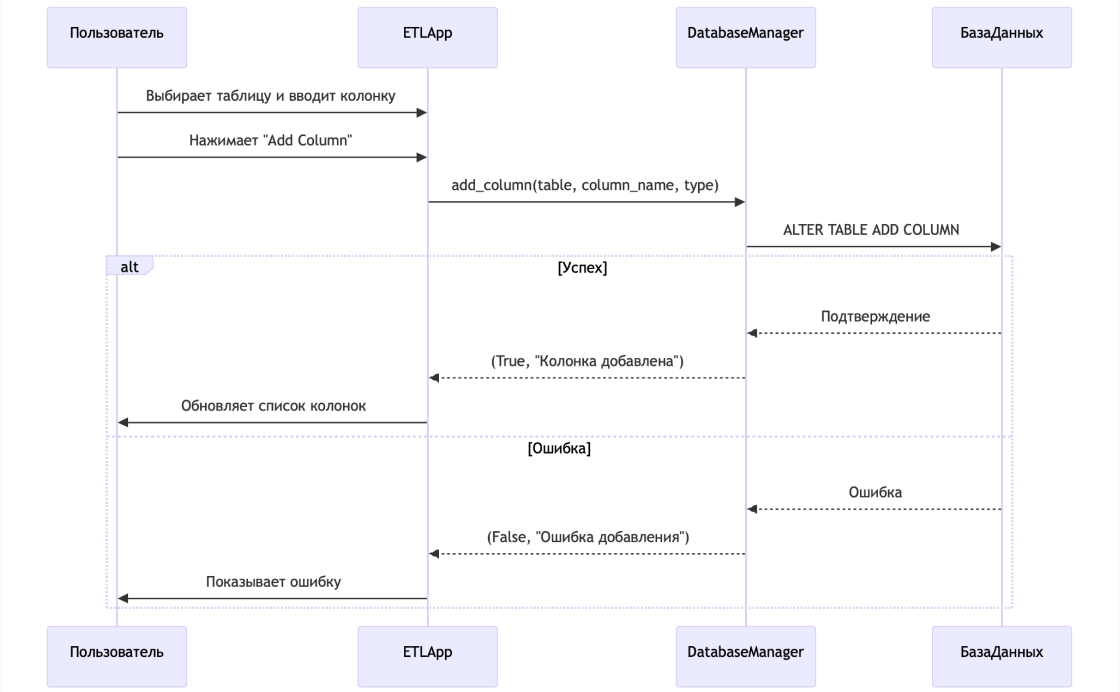


Диаграмма последовательности "Выполнение ETL-процесса"

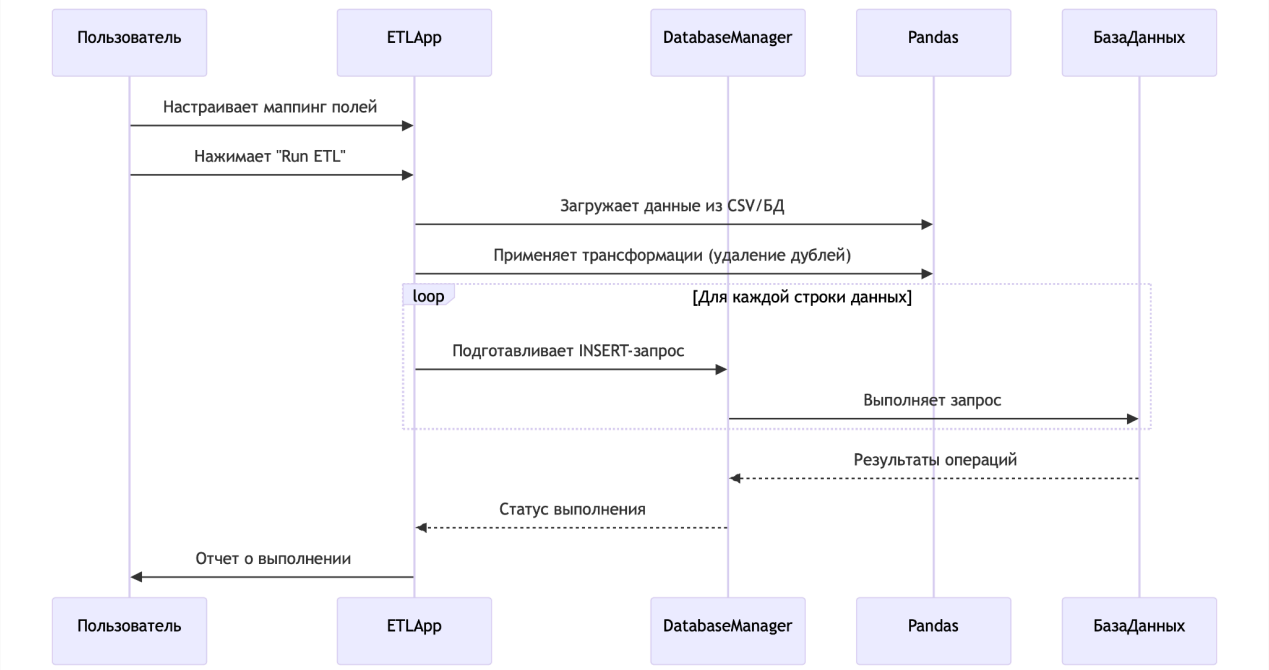
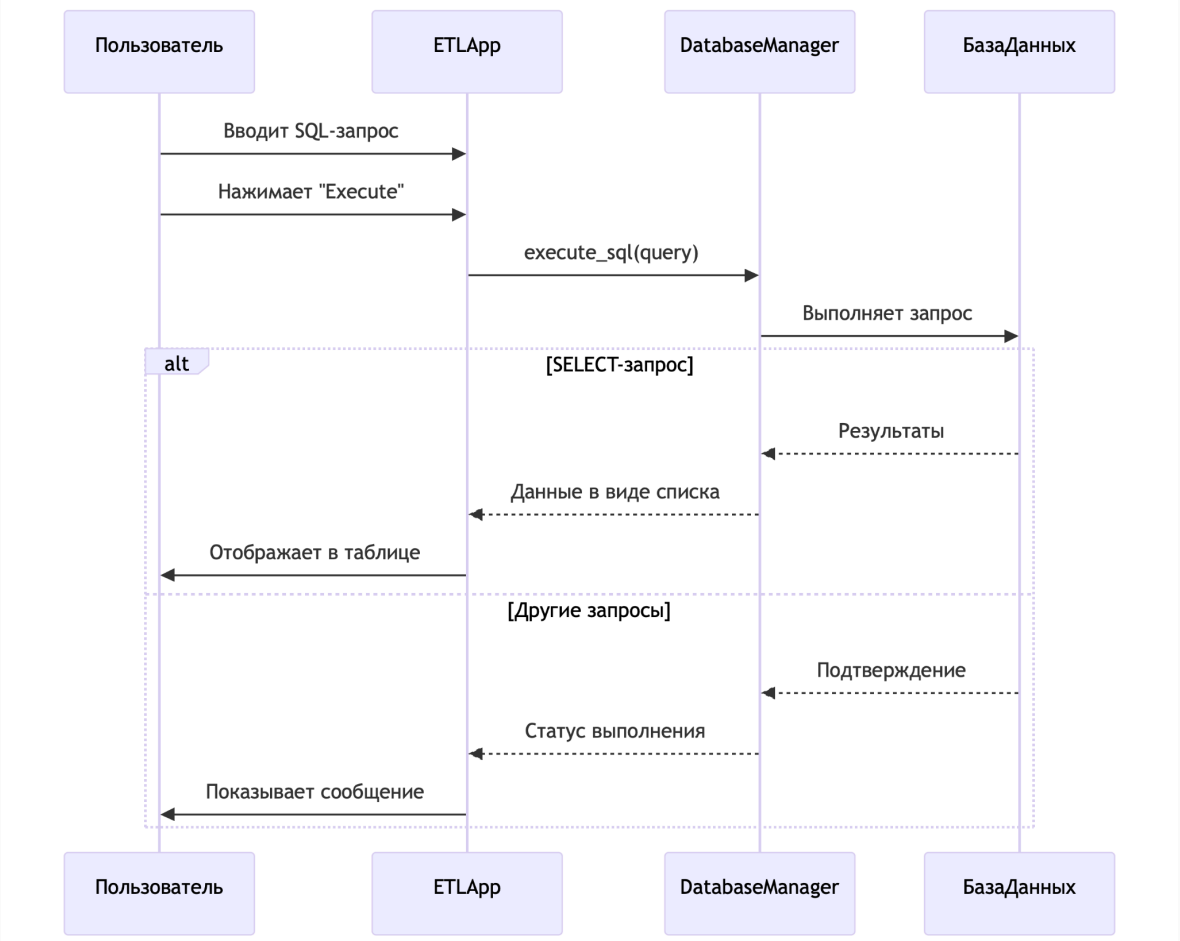
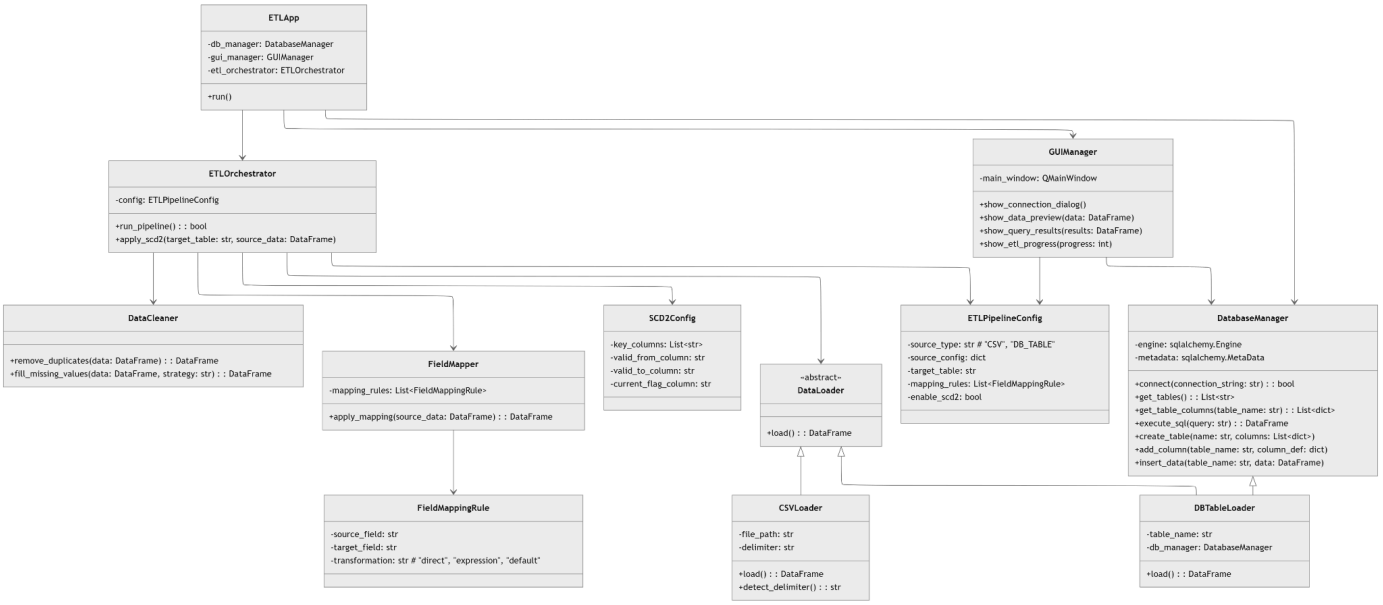


Диаграмма последовательности "Выполнение SQL-запроса"



2.4 Диаграмма проектных классов (UML)



3. Этап реализации

Полный листинг кода представлен в предыдущем сообщении. Основные компоненты:

3.1 Архитектура

Приложение использует трехуровневую архитектуру:

1. **Презентационный уровень**: PyQt5 интерфейс
2. **Бизнес-логика**: Класс ETLApp
3. **Уровень данных**: DatabaseManager и SQLAlchemy

3.2 Ключевые классы

1. **DatabaseManager**: Инкапсулирует работу с БД
2. **ETLApp**: Основной класс приложения, реализует GUI и бизнес-логику

3.3 Основные методы

- `connect_to_db()`: Установка соединения с БД
- `load_csv_file()`: Загрузка данных из CSV
- `remove_duplicates()`: Очистка данных
- `create_table()`: Создание таблицы в БД
- `run_etl_process()`: Выполнение ETL процесса
- `execute_query()`: Выполнение SQL запроса

4. Этап тестирования

4.1 План тестирования

1. **Тестирование подключения к БД**:
    - Корректные учетные данные
    - Неверные учетные данные
    - Недоступный сервер
2. **Тестирование загрузки данных**:
    - Корректный CSV
    - Поврежденный CSV
    - Загрузка из существующей таблицы
    - Загрузка из несуществующей таблицы
3. **Тестирование очистки данных**:
    - Данные с дубликатами
    - Данные без дубликатов
    - Пустые данные
4. **Тестирование работы с таблицами**:
    - Создание таблицы
    - Добавление колонки
    - Попытка создания существующей таблицы
5. **Тестирование ETL процесса**:
    - Простой маппинг полей
    - Маппинг с преобразованием типов
    - SCD2 процесс

6. **Тестирование запросов**:
    - SELECT запросы
    - INSERT/UPDATE запросы
    - Невалидные SQL запросы

4.2 Протокол тестирования (скриншоты)

Примечание: В реальном отчете здесь должны быть скриншоты:

1. **Успешное подключение к БД**:
    - Заполненные поля подключения
    - Сообщение об успешном подключении
2. **Загрузка CSV файла**:
    - Выбор файла через диалог
    - Отображение превью данных
3. **Очистка данных**:
    - Данные до и после очистки
    - Сообщение о количестве удаленных дубликатов
4. **Создание таблицы**:
    - Форма с названием и колонками
    - Сообщение об успешном создании
5. **Настройка маппинга**:
    - Выбор полей источника и назначения
    - Таблица с настроенными маппингами
6. **Выполнение ETL**:
    - Настройки SCD2
    - Сообщение об успешном завершении
7. **Выполнение запроса**:
    - Введенный SQL запрос
    - Результаты в табличном виде

5. Заключение

Разработанное приложение удовлетворяет всем поставленным требованиям:

- Реализован графический интерфейс для управления ETL процессами
- Поддерживаются основные источники данных (CSV, PostgreSQL, MySQL)
- Обеспечена возможность очистки данных и управления структурой БД
- Реализована поддержка SCD2 для историчности данных
- Предоставлен интерфейс для выполнения SQL запросов

Приложение может быть расширено за счет:

- Добавления поддержки других СУБД
- Реализации более сложных преобразований данных
- Добавления планировщика ETL-заданий
- Экспорта/импорта конфигураций ETL процессов

Ссылки:

https://colab.research.google.com/drive/1O-qaO34ieKFcub_MN-ff7Ru8A-Uk3Z7O#scrollTo=vhSpsGBEC1YD - код

https://docs.google.com/document/d/1jp3ZQX7x0336w5yuWcW17CVKxEc8al8tpllLuQ8YoE0/edit?tab=t.0#heading=h.tv6dz49qzy6s - спека

https://docs.google.com/document/d/1k6sL8htLHA1j4LPB3cpe2urd2JXW0tJE0WIlHajIfVE/edit?usp=sharing - план тестирования

https://docs.google.com/document/d/1FQ9IiHQqrrMAk7Y4lpQ93mwa6BVSKXyKkNDYZyenXnU/edit?tab=t.0#heading=h.wal644d3blrd - руководство пользователя