

# Руководство пользователя: ETL Flow Manager

## 1. Введение

**ETL Flow Manager** — это графическое приложение для автоматизации процессов **Extract, Transform, Load (ETL)**.

Оно позволяет:

- Подключаться к базам данных (**PostgreSQL, MySQL**).
- Загружать данные из **CSV-файлов** и таблиц БД.
- Очищать и преобразовывать данные.
- Создавать и изменять структуру таблиц.
- Настраивать **SCD2 (медленно меняющиеся измерения)**.
- Выполнять **SQL-запросы** и экспортировать результаты.

### Целевая аудитория:

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

## 2. Установка

### 2.1. Системные требования

- **ОС:** Windows 10+, Ubuntu 20.04+, macOS 12+
- **Python:** 3.8+
- **Зависимости:**

```
pip install pandas sqlalchemy PyQt5
```

## 2.2. Запуск приложения

1. Скачайте проект с репозитория.
2. Перейдите в папку с проектом:

```
cd /path/to/etl_flow_manager
```

3. Запустите приложение:

```
python main.py
```

## 3. Основные функции

### 3.1. Подключение к базе данных

1. Перейдите во вкладку "**Database Connection**".
2. Выберите тип БД (**PostgreSQL/MySQL**).
3. Введите параметры подключения:
  - **Host** (хост)
  - **Port** (порт)
  - **Database Name** (имя базы)
  - **Username** (логин)
  - **Password** (пароль)
4. Нажмите "**Connect**".
5. **Статус подключения** отобразится внизу окна.

#### Пример успешного подключения:

```
Connected to PostgreSQL: my_database
```

#### Пример ошибки:

```
Connection failed: Invalid credentials
```

## 3.2. Загрузка данных

### Из CSV-файла

1. Перейдите во вкладку **"Data Source"**.
2. Нажмите **"Select CSV File"**.
3. Выберите файл (например, `sample_data/customers.csv`).
4. Данные отобразятся в таблице предпросмотра.

### Из таблицы БД

1. Убедитесь, что подключение к БД активно.
2. Выберите таблицу из выпадающего списка.
3. Нажмите **"Load Table Data"**.
4. Данные загрузятся в интерфейс.

## 3.3. Очистка данных

1. Перейдите во вкладку **"Data Cleaning"**.
2. Нажмите **"Remove Duplicates"**.
3. Приложение покажет количество удаленных дубликатов:

`Removed 5 duplicate rows.`

### 3.4. Создание и изменение таблиц

#### Создание новой таблицы

1. Перейдите во вкладку **"Table Creation"**.
2. Введите **имя таблицы** (например, `customers`).
3. Добавьте колонки:
  - **Имя колонки** (например, `customer_id`).
  - **Тип данных** (например, `INTEGER`, `VARCHAR(50)`).
  - **Primary Key** (отметьте, если нужно).
  - **Nullable** (разрешить `NULL`).
4. Нажмите **"Create Table"**.

#### Пример успешного создания:

Table "customers" created successfully.

#### Добавление колонки

1. Выберите существующую таблицу.
2. Нажмите **"Add Column"**.
3. Введите параметры новой колонки.
4. Подтвердите действие.

### 3.5. Настройка ETL-процесса (Field Mapping)

1. Перейдите во вкладку **"Field Mapping"**.
2. Выберите **источник данных** (CSV или таблица БД).
3. Выберите **целевую таблицу**.
4. Настройте соответствие полей:
  - `source_field` → `target_field` (например, `customer_id` → `client_id`).
5. Для **SCD2**:
  - Включите чекбокс **"Enable SCD2"**.
  - Укажите названия полей:
    - `valid_from` (дата начала актуальности).
    - `valid_to` (дата окончания).
    - `is_current` (флаг актуальности).
6. Нажмите **"Run ETL"**.

#### Пример успешного выполнения:

ETL process completed. 100 rows inserted.

### 3.6. Выполнение SQL-запросов

1. Перейдите во вкладку **"Query Interface"**.
2. Введите SQL-запрос (например, `SELECT * FROM customers;`).
3. Нажмите **"Execute Query"**.
4. Результаты отобразятся в таблице.
5. Для экспорта в CSV нажмите **"Export to CSV"**.

## 4. Примеры использования

### 4.1. Миграция данных из CSV в PostgreSQL

1. Подключитесь к PostgreSQL.
2. Загрузите CSV (customers.csv).
3. Создайте таблицу customers с нужной структурой.
4. Настройте маппинг полей.
5. Запустите ETL-процесс.

### 4.2. Реализация SCD2

1. Создайте таблицу customers\_history с полями:
  - customer\_id (PK)
  - name
  - valid\_from (DATE)
  - valid\_to (DATE)
  - is\_current (BOOLEAN)
2. Загрузите исходные данные.
3. Включите SCD2 в настройках маппинга.
4. Запустите ETL.

## 5. Обработка ошибок

Ошибка	Решение
Connection failed	Проверьте логин/пароль и доступность БД.
Failed to load CSV	Убедитесь, что файл существует и имеет правильный формат.
No field mappings	Настройте соответствие полей перед запуском ETL.
Invalid SQL syntax	Проверьте запрос на ошибки.

## 6. Заключение

**ETL Flow Manager** упрощает работу с ETL-процессами благодаря:

- **Графическому интерфейсу** (PyQt5).
- **Поддержке PostgreSQL и MySQL.**
- **Гибкой настройке маппинга полей.**
- **SCD2 для хранения исторических данных.**

Для расширения функциональности можно добавить:

- **Поддержку других СУБД** (SQLite, Oracle).
- **Планировщик ETL-заданий.**
- **Импорт/экспорт конфигураций.**