Отчет: Загрузчик данных для стажировки Simbirsoft

1. Введение

Python-приложение для загрузки данных, разработанное в рамках тестового задания на стажировку Backend в Simbirsoft. Приложение загружает данные из CSV-файла, доступного по предоставленному URL-адресу, обрабатывает их с помощью библиотеки Pandas и загружает в базу данных PostgreSQL. Все решение контейнеризировано с использованием Docker.

2. Функциональное описание

Приложение выполняет следующие основные функции:

- **Получение данных:** Загружает CSV-данные с URL-адреса, предоставленного пользователем.
- **Трансформация данных:**
    - Использует Pandas для чтения и обработки данных CSV.
    - Очищает и форматирует имена столбцов для совместимости с базой данных.
    - Определяет подходящие типы данных для каждого столбца.
- **Загрузка данных:**
    - Создает таблицу на основе структуры CSV-файла.
    - Загружает данные в базу данных с помощью команды COPY.
- **Журналирование:** Записывает события приложения и ошибки в консоль и файл журнала.
- **Обработка ошибок:** При возникновении ошибок программа прекращает свою работу, база данных откатывается до исходного состояния.

3. Архитектура приложения

Приложение имеет модульную структуру, где каждый компонент отвечает за определенные задачи:

- **main.py:** Управляет процессом загрузки данных.
- **loader.py:** Содержит функции для проверки доступности URL и загрузки данных.
- **database.py:** Отвечает за подключение к базе данных, создание таблиц и загрузку данных.
- **csv_functions.py:** Предоставляет утилиты для обработки данных CSV.
- **logger.py:** Настраивает систему журналирования.

4. Схема базы данных

Таблицы создаются динамически на основе структуры полученного CSV-файла. Имя таблицы и типы колонок извлекаются из заголовков и данных CSV-файлов. Первичные ключи явно не определены.

5. Техническая реализация

5.1 Инструменты разработки и библиотеки

- **Язык:** Python 3.11
- **Библиотеки:**
  - requests: Получение данных с URL-адреса.
  - psycopg2-binary: Взаимодействие с базой данных PostgreSQL.
  - pandas: Анализ и обработка данных.
  - python-dotenv: Загрузка переменных окружения из файла .env.
- **Менеджер пакетов:** Poetry
- **Контейнеризация:** Docker

6. Конфигурация Docker

- **Dockerfile:**
  - Создает образ приложения Python.
  - Устанавливает зависимости с помощью Poetry.
- **docker-compose.yml:**
  - Определяет два сервиса:
    - **postgres:** Запускает контейнер базы данных PostgreSQL.
    - **python:** Запускает контейнер приложения Python и связывает его с контейнером базы данных.

7. Развертывание и запуск

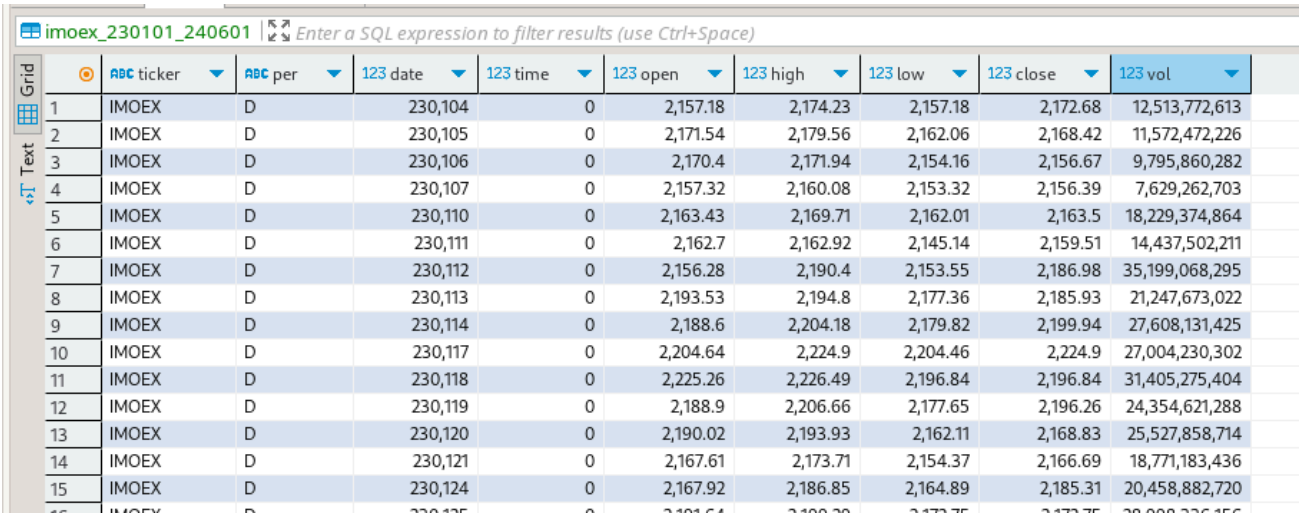
- `git clone https://github.com/Orlovchikk/simbirsoft-internship.git`
- `cd simbirsoft-internship`
- создать файл `.env` и заполнить по образцу `.env.example`
- `docker compose up -d`
- `docker compose exec python bash`
- `poetry run python main.py '**link**'`

8. Скриншоты, демонстрирующие работу приложения

Вызов команды `poetry run python main.py https://drive.usercontent.google.com/download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq&export=download`

```
root@1d47a45d0684:/app# poetry run python main.py https://drive.usercontent.google.com/download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq&export=download
01/08/2024 08:01:40 - INFO: Starting data loading process for URL: https://drive.usercontent.google.com/download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq
01/08/2024 08:01:40 - INFO: Attempting to fetch data from: https://drive.usercontent.google.com/download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq
01/08/2024 08:01:47 - INFO: Data fetched successfully
01/08/2024 08:01:47 - INFO: Extracted table name: IMOEX_230101_240601
01/08/2024 08:01:47 - INFO: Converting data to DataFrame
01/08/2024 08:01:47 - INFO: Data converted successfully
01/08/2024 08:01:47 - INFO: Connecting to database: postgres
01/08/2024 08:01:47 - INFO: Connected successfully
01/08/2024 08:01:47 - INFO: Starting data loading process to table: imoex_230101_240601
01/08/2024 08:01:47 - INFO: Data successfully uploaded to table 'imoex_230101_240601'
01/08/2024 08:01:47 - INFO: Data loading process completed.
```

Созданная таблица `imoex_230101_240601`. Просмотр в приложении DBeaver (для входа используются данные из файла .env)



imoex_230101_240601 | Enter a SQL expression to filter results (use Ctrl+Space)

Grid | Text

	ticker	per	date	time	open	high	low	close	vol
1	IMOEX	D	230,104	0	2,157.18	2,174.23	2,157.18	2,172.68	12,513,772,613
2	IMOEX	D	230,105	0	2,171.54	2,179.56	2,162.06	2,168.42	11,572,472,226
3	IMOEX	D	230,106	0	2,170.4	2,171.94	2,154.16	2,156.67	9,795,860,282
4	IMOEX	D	230,107	0	2,157.32	2,160.08	2,153.32	2,156.39	7,629,262,703
5	IMOEX	D	230,110	0	2,163.43	2,169.71	2,162.01	2,163.5	18,229,374,864
6	IMOEX	D	230,111	0	2,162.7	2,162.92	2,145.14	2,159.51	14,437,502,211
7	IMOEX	D	230,112	0	2,156.28	2,190.4	2,153.55	2,186.98	35,199,068,295
8	IMOEX	D	230,113	0	2,193.53	2,194.8	2,177.36	2,185.93	21,247,673,022
9	IMOEX	D	230,114	0	2,188.6	2,204.18	2,179.82	2,199.94	27,608,131,425
10	IMOEX	D	230,117	0	2,204.64	2,224.9	2,204.46	2,224.9	27,004,230,302
11	IMOEX	D	230,118	0	2,225.26	2,226.49	2,196.84	2,196.84	31,405,275,404
12	IMOEX	D	230,119	0	2,188.9	2,206.66	2,177.65	2,196.26	24,354,621,288
13	IMOEX	D	230,120	0	2,190.02	2,193.93	2,162.11	2,168.83	25,527,858,714
14	IMOEX	D	230,121	0	2,167.61	2,173.71	2,154.37	2,166.69	18,771,183,436
15	IMOEX	D	230,124	0	2,167.92	2,186.85	2,164.89	2,185.31	20,458,882,720

Обработка ошибки “Invalid URL”

```
root@1d47a45d0684:/app# poetry run python main.py asdfasa
Skipping virtualenv creation, as specified in config file.
01/08/2024 08:12:39 - INFO: Starting data loading process for URL: asdfasa
01/08/2024 08:12:39 - INFO: Attempting to fetch data from: asdfasa
01/08/2024 08:12:39 - ERROR: URL is not downloadable 'asdfasa': Invalid URL 'asdfasa': No scheme supplied. Perhaps you meant https://asdfasa?
01/08/2024 08:12:39 - ERROR: Failed to fetch data. Exiting.
```

Первые строчки из файла python-script.log

```
root@1d47a45d0684:/app/app/logs# head python_script.log
[INFO|main|L11] 01/08/2024 08:00:29: Starting data loading process for URL: https://drive.usercontent.google.com/download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq
[INFO|loader|L23] 01/08/2024 08:00:29: Attempting to fetch data from: https://drive.usercontent.google.com/download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq
[DEBUG|loader|L7] 01/08/2024 08:00:29: Checking if URL is downloadable: https://drive.usercontent.google.com/download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq
[DEBUG|connectionpool|L1051] 01/08/2024 08:00:29: Starting new HTTPS connection (1): drive.usercontent.google.com:443
[DEBUG|connectionpool|L546] 01/08/2024 08:00:34: https://drive.usercontent.google.com:443 "HEAD /download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq HTTP/11" 200 0
[DEBUG|loader|L14] 01/08/2024 08:00:34: URL is downloadable
[DEBUG|connectionpool|L1051] 01/08/2024 08:00:34: Starting new HTTPS connection (1): drive.usercontent.google.com:443
[DEBUG|connectionpool|L546] 01/08/2024 08:00:40: https://drive.usercontent.google.com:443 "GET /download?id=1RJKGrwzznR9wA7rGNM76cRp1Tis7V-Xq HTTP/11" 200 23937
[INFO|loader|L30] 01/08/2024 08:00:40: Data fetched successfully
[DEBUG|api|L439] 01/08/2024 08:00:41: Encoding detection: ascii is most likely the one.
```