ETL File Description

Introduction

The project is implemented using the Python language. The psycopg2 library is required for connecting to the database.

Initially, install the psycopg2 library using the following command:

pip install psycopg2

The ETL folder contains three Python files:

- create_library.py: This script creates the database schema for the library database.
- create_warehouse.py: This script creates the schema for the data warehouse.

Both scripts use the SCHEMA and DATA MODEL files designed based on the given requirements. The design of this part was inspired by a similar code on GitHub.

The second phase logic is executed by the ETL.py file.

Explanation of the ETL.py File

The ETL.py file performs the ETL operations. Here's a brief explanation of its structure and functions:

1. Delete Operation:
   - If data is deleted from table X in the source database, the corresponding records are deleted from the DELETE_X table in the destination database.
   - The records are transferred based on the DELETE operation and their PRIMARY KEY is stored in the data warehouse.

2. Update Operation:

- If data is updated in table X in the source database, the corresponding records are updated in the UPDATE_X table in the destination database.

- The records are transferred based on the UPDATE operation and their PRIMARY KEY is stored in the data warehouse.

3. Insert Operation:

- If new data is inserted into table X in the source database, the corresponding records are inserted into the INSERT_X table in the destination database.

- The records are transferred based on the CREATE operation.

Connecting to the Database in ETL.py

To connect to the database, the following lines of code are used at the beginning of the file:

```
import psycopg2

conn_source = psycopg2.connect('dbname=source_db user=your_username password=your_password')
conn_warehouse = psycopg2.connect('dbname=warehouse_db user=your_username password=your_password')
```

Explanation of the REFACTOR Function

The REFACTOR function is used to convert the data file to a suitable format for the database. This function replaces all None values in the data with empty strings and converts them to their string representation.