Dokumentasi Simple ETL Dengan CSV ke PostgreSQL
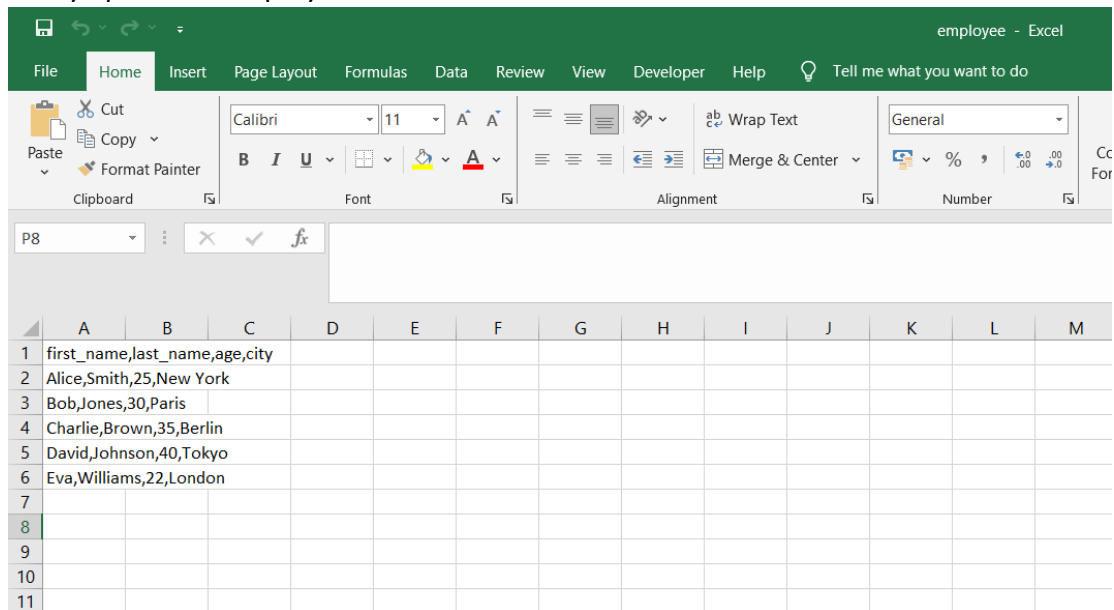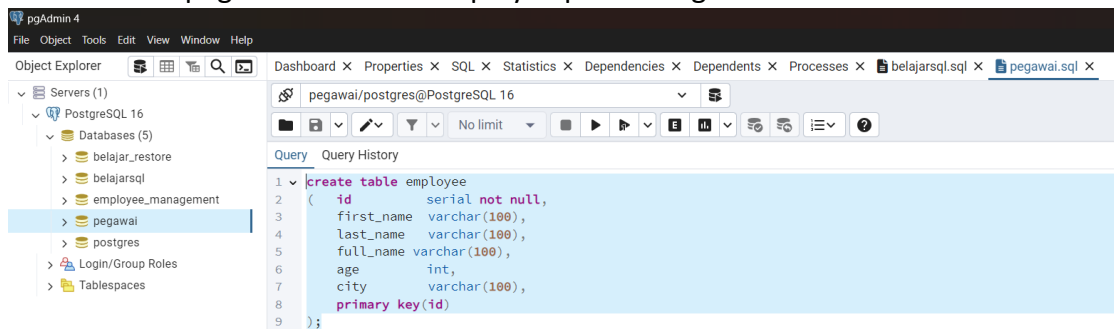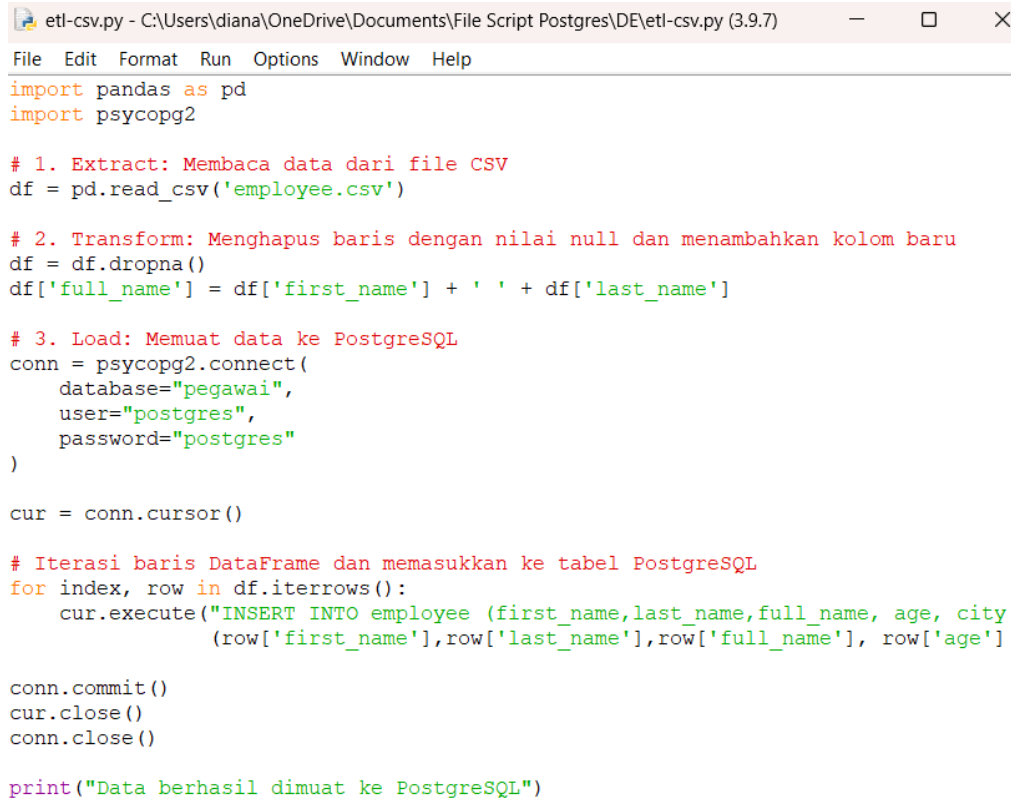
1. Rancangan sistem ETL
   a. Read data csv
   b. Transform data (hapus null jika ada dan menambahkan full_name)
   c. Load data ke PostgreSQL
2. Menyiapkan file employee.csv untuk di ekstrak

3. Membuat db pegawai dan tabel employee pada PostgreSQL

4. Menyiapkan file ETL dengan menggunakan python kemudian merunningnya

etl-csv.py - C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py (3.9.7) — □ ✕

File   Edit   Format   Run   Options   Window   Help

```python
import pandas as pd
import psycopg2

# 1. Extract: Membaca data dari file CSV
df = pd.read_csv('employee.csv')

# 2. Transform: Menghapus baris dengan nilai null dan menambahkan kolom baru
df = df.dropna()
df['full_name'] = df['first_name'] + ' ' + df['last_name']

# 3. Load: Memuat data ke PostgreSQL
conn = psycopg2.connect(
    database="pegawai",
    user="postgres",
    password="postgres"
)

cur = conn.cursor()

# Iterasi baris DataFrame dan memasukkan ke tabel PostgreSQL
for index, row in df.iterrows():
    cur.execute("INSERT INTO employee (first_name,last_name,full_name, age, city
                (row['first_name'],row['last_name'],row['full_name'], row['age']

conn.commit()
cur.close()
conn.close()

print("Data berhasil dimuat ke PostgreSQL")
```

```
IDLE Shell 3.9.7                                            —   □   ×
File   Edit   Shell   Debug   Options   Window   Help
                                   ^
>>>
= RESTART: C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py
Traceback (most recent call last):
  File "C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py", l
ine 22, in <module>
    cur.execute("INSERT INTO employee (full_name, age, city) VALUES (%s, %s, %s)
",
psycopg2.errors.UndefinedColumn: column "full_name" of relation "employee" does
not exist
LINE 1: INSERT INTO employee (full_name, age, city) VALUES ('Alice S...
                                   ^

>>>
= RESTART: C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py
Traceback (most recent call last):
  File "C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py", l
ine 22, in <module>
    cur.execute("INSERT INTO employee (full_name, age, city) VALUES (%s, %s, %s)
",
psycopg2.errors.NotNullViolation: null value in column "first_name" of relation
"employee" violates not-null constraint
DETAIL:  Failing row contains (1, null, null, 25, New York, Alice Smith).

>>>
= RESTART: C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py
Data berhasil dimuat ke PostgreSQL
>>>
= RESTART: C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py
Traceback (most recent call last):
  File "C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py", l
ine 22, in <module>
    cur.execute("INSERT INTO employee (first_name,last_name,full_name, age, city
) VALUES (%s, %s, %s)",
TypeError: not all arguments converted during string formatting
>>>
= RESTART: C:\Users\diana\OneDrive\Documents\File Script Postgres\DE\etl-csv.py
Data berhasil dimuat ke PostgreSQL
>>> |
                                                        Ln: 50   Col: 4
```

5. Data berhasil di load kedalam db pegawai