Movies-ETL

Project Overview

Amazing Prime loves the dataset and wants to keep it updated on a daily basis. They would like to create an automated pipeline that takes in new data, performs the appropriate transformations, and loads the data into existing tables.

Purpose: -

- To write ETL function to read Wikipedia data, Kaggle metadata, and the MovieLens rating data and to create three separate DataFrames.
- To extract and transform the Wikipedia and Kaggle data
- To create the movie database; PostgresSQL movie Database

Resources: -

- wikipedia-movies.json, movies_metadata.csv and ratings.csv

Results: - (please refer to the attached files for details)

- **Deliverable 1**: ETL Function was written to read three data files
- **Deliverable 2 and 3:** Wikipedia and Kaggle data were extracted and transformed
- **Deliverable 4:** Movie database was created

Load the data to a PostgregSQL movie database

```
  db_string = f"postgresql://postgres:{db_password}@localhost:5432/movie_data"
    # create database engine
    engine = create_engine(db_string)
    movies_df.to_sql(name='movies', con=engine, if_exists='replace')
    rows_imported = 0
# get the start_time from time.time()
    start_time = time.time()
    for data in pd.read_csv(f'ratings.csv', chunksize=1000000):
        print(f'importing rows {rows_imported} to {rows_imported + len(data)}...', end='')
        data.to_sql(name='ratings', con=engine, if_exists='append')
        rows_imported += len(data)
    # add elapsed time to final print out
        print(f'Done. {time.time() - start_time} total seconds elapsed')
```

```
In [5]: extract_transform_load()

video
importing rows 0 to 1000000...Done. 84.43873023986816 total seconds elapsed
importing rows 1000000 to 2000000...Done. 174.6356692314148 total seconds elapsed
importing rows 2000000 to 3000000...Done. 265.2194800376892 total seconds elapsed
importing rows 3000000 to 4000000...Done. 350.21102809906006 total seconds elapsed
importing rows 4000000 to 5000000...Done. 442.8558871746063 total seconds elapsed
importing rows 5000000 to 6000000...Done. 527.345939874649 total seconds elapsed
importing rows 6000000 to 7000000...Done. 611.5172863006592 total seconds elapsed
importing rows 7000000 to 8000000...Done. 693.6667623519897 total seconds elapsed
importing rows 8000000 to 9000000...Done. 775.0011265277863 total seconds elapsed
importing rows 9000000 to 10000000...Done. 860.9964861869812 total seconds elapsed
importing rows 10000000 to 11000000...Done. 944.1840279102325 total seconds elapsed
importing rows 11000000 to 12000000...Done. 1028.1161913871765 total seconds elapsed
importing rows 12000000 to 13000000...Done. 1110.8113751411438 total seconds elapsed
importing rows 13000000 to 14000000...Done. 1199.5469632148743 total seconds elapsed
importing rows 14000000 to 15000000...Done. 1291.8761949539185 total seconds elapsed
importing rows 15000000 to 16000000...Done. 1373.5205295085907 total seconds elapsed
importing rows 16000000 to 17000000...Done. 1461.046995639801 total seconds elapsed
importing rows 17000000 to 18000000...Done. 1545.8920304775238 total seconds elapsed
importing rows 18000000 to 19000000...Done. 1630.552181005478 total seconds elapsed
importing rows 19000000 to 20000000...Done. 1726.2988874912262 total seconds elapsed
importing rows 20000000 to 21000000...Done. 1816.1384437084198 total seconds elapsed
importing rows 21000000 to 22000000...Done. 1900.016221523285 total seconds elapsed
importing rows 22000000 to 23000000...Done. 1972.317176580429 total seconds elapsed
importing rows 23000000 to 24000000...Done. 2045.4673628807068 total seconds elapsed
importing rows 24000000 to 25000000...Done. 2118.5832369327545 total seconds elapsed
importing rows 25000000 to 26000000...Done. 2190.8187053203583 total seconds elapsed
importing rows 26000000 to 26024289...Done. 2192.5299112796783 total seconds elapsed
```