Lab2_DW_ETL_Report

Group members: Sehrish Naqvi, Waleed Tariq, Balaji Vijayraj, Charu Bisht

In this assignment, we are implementing an ETL in Python, using the Pandas library. Here we are going to extract data using The Movie Database API.

EXTRACT:

We can extract data from different sources. We extracted data from "The Movie Database". For extracting the data from the database, we obtained an API key with the help of the website mentioned and used the get method to request 6 movies with ratings from 550 to 555. We formatted the data from JSON responses to the panda's data frame. The data set has 6 rows and 25 columns.

TRANSFORM:

We created a list of columns from the main data frame as mentioned and have it as a new data frame called **df_bck**. We added another column **genre_name** in which we combined all the values from the genre column which were originally in the 'list of dictionaries' formats through a function. We put those in a separate array. For the second function, we used a newly created column and created a separate array called **total_genre**. We compared the list with **total_genre**. If they are the same, then the frequency is 1, else it is 0. We do this for each row of the genre and created the **genre_freq** column. We then split the column in the same sequence as **total_genre** so that we have a column for each categorical value. For this, we created a separate table for the genre and a column of list to explode out (explode ().to_list() method). We created a datetimes data frame with **release_date** column and then converted it into datetime format.

LOAD:

We have exported 3 tables called 'movies', 'genres' and 'datetimes' to a csv file.

For VG

As per the instructions, we have transformed the runtime in hours and minutes. For loading data into SQL lite relational database, we established a connection to the database, then loaded data frames to the database. For verification, we have also retrieved the data back from the database.