Northeastern

Final Project Document
by
(Group- 8)
Course: INFO 7290 Data Warehouse & Business Intelligence SEC 01

Group Members:
Abhishek Shetty
Kumaran Nehru Uthra
Sai Sirisha Pyarasani
Soumya Gummalla

Document Revision History:

Date	Version	Changes	Author
11/5/2020	1.0	Initial document	Team
11/15/2020	1.1	Added additional details regarding data source, described the datasets and added professors' comments to end of document as an appendix	Team
12/04/2020	2	Added awards data, removed professor's questions from the document, Updated error handling and data logging, Loaded data in SSMS, added screenshot of SSIS packages and analysed data	Team
12/16/2020	3	Added automated validations (Logging), gave details of how the data has been scrapped, gave details of how the star schema has been built, built an OLAP cube, gave examples of row count, error handling, SCD	Team

Table of Contents

1.Introduction:

Movies have a significant impact on mankind as they are a combination of audio, video, voice, and graphics. Movies help us understand the world we live in, from different points of view. Success of a movie depends on the degree to which the viewer mentally connects to it, but the commercial success of a movie depends on the director, the actors, and the genre. In this project we aim at understanding what makes a movie more successful compared to the others.

2.Objectives:

- To get data from different data sources.
- To build a dimensional model from the Imdb movies data.
- To automate ETL to load and combine datasets using SSIS packages.
- To process the data and obtain an OLAP data cube.
- To analyse the data using visualizations.

3.Data

The data has been denormalised into different dimensions to build the snowflake schema. The data includes 3 main dimensions, which are:

- **IMDb Movies:**

Movie data has further been divided into reviews and crew data. This Data source file contain the details like:

Imdb_title_id: This column contains the Movie title IDs as given on IMDb
title: This column holds the name of the movies. It contains 82094 unique string values.
original_title: This column contains the original tile name of the movie. It contains 80852 unique values of the String data type.
year: This column holds integer type data. It contains information about the Year when the movie was released.
genre: This column contains data about the genre of the movie. There are 1257 unique genres.
country: This column contains data about the name of the countries this movie was released in. String data type is stored in this column.
language: This column holds the data about the languages this movie is available in. String data type is stored in this column.
director : This column contains the name of the movie director. It contains 87 missing records. String data type is stored in this column.
writer : This column contains the name of the movie writers. It contains 1672 missing records. String data type is stored in this column.
production_company : This column contains the name of the movie production company. It contains 4455 missing records. String data type is stored in this column.
budget : budget column gives us data about the total budget of the movie. The currency in which the budget is mentioned needs to be standardised. This column contains data of the String type. It contains 62100 missing records.
usa_gross_income : usa_gross_income column gives us data about the gross income of the movie in the United States. This column contains data of the String type. It contains 70500 missing records.

worlwide_gross_income : worldwide_gross_income column gives us data about the gross income of the movie worldwide. This column contains data of the String type.It contains 54800 missing records.

- **IMDb Ratings Data:**

This data source contains the data related to rating given to the movie with respect to different demographic groups. Based on the demographics the Rating table has been split into 2 types which are based on gender and geography, respectively. The data contains details like:

total votes received: This column contains the sum of votes received to rate the particular movie.

total median vote: This column gives the median values of ratings from all the votes received by the movie

number of votes with rating equal to 10: This column gives the total number of votes that gave a rating of 10 to that particular movie.

average rating from all male users with demographic data available (all ages): This column contains the mean of rating given by all male users irrespective of their age.

average rating from all female users with demographic data available (all ages): This column contains the count of votes for a particular movie given by all female users irrespective of their age.

number of votes from US voters: This column contains the count of votes given by voters who belong to the US

number of votes from non-US voters: This column contains the count of votes given by voters who do not belong to the US

- **Awards Data:**

This data set contains the details of Academy awards that have been received by the movies.

Its attributes are:

Film: Title of the film.

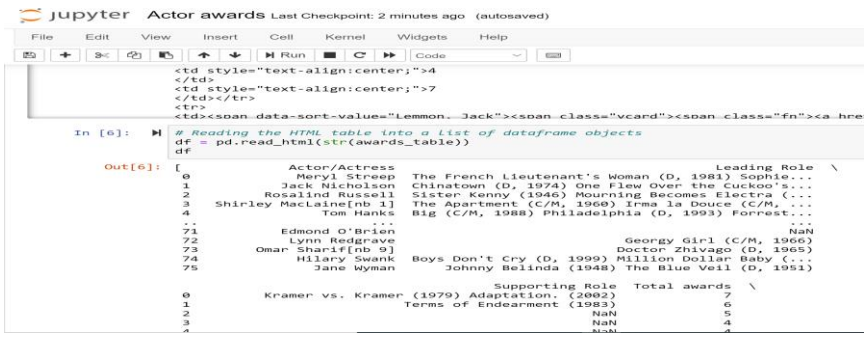
Awards: Number of awards received by the film.

Nominations: Number of times the movie has been nominated for the awards.

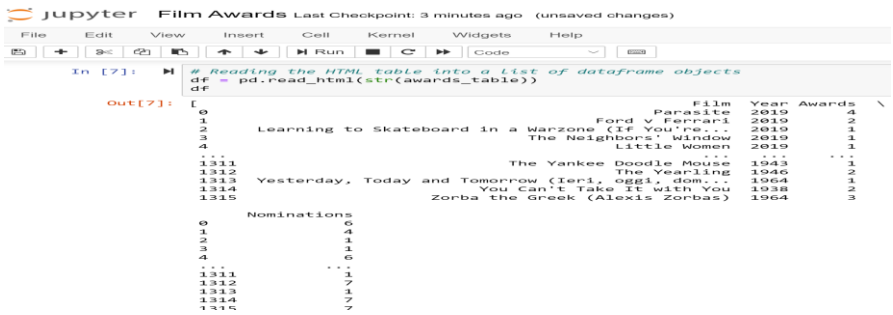
Web Scraping:

The awards data has been scraped from wikipedia using python as shown below.

```
In [1]:
# Importing libraries
from bs4 import BeautifulSoup
import requests
import pandas as pd
from pandas import DataFrame

In [2]:
# URL from which we are going to scrap the data
wiki_url = 'https://en.wikipedia.org/wiki/Golden_Globe_Awards'
# Extracting the specific tag for webscraping
tableclass = "wikitable sortable"

In [12]:
#Checking the connecting of the URL
response = requests.get(wiki_url)
response

Out[12]: <Response [200]>

In [4]:
# Parsing the webpage
soup = BeautifulSoup(response.text,'html.parser')
soup
```



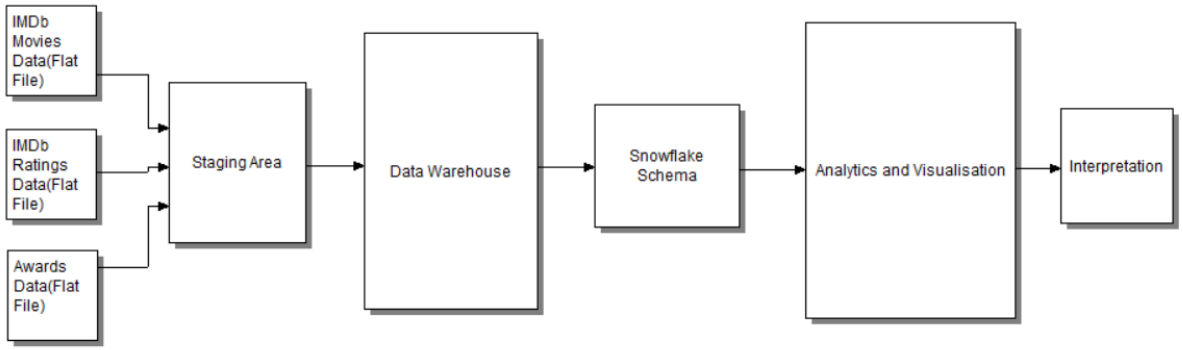
```
jupyter Actor awards Last Checkpoint: 2 minutes ago (autosaved)
File Edit View Insert Cell Kernel Widgets Help
Run Code

<td style="text-align:center;">4
</td>
<td style="text-align:center;">7
</td></tr>
<tr>
<td><span data-sort-value="Lemmon, Jack"><span class="vcard"><span class="fn"><a hre

In [6]:
# Reading the HTML table into a list of dataframe objects
df = pd.read_html(str(awards_table))
df

Out[6]: [            Actor/Actress                                       Leading Role  \
0          Meryl Streep  The French Lieutenant's Woman (D, 1981) Sophie...
1         Jack Nicholson  Chinatown (D, 1974) One Flew Over the Cuckoo's...
2        Rosalind Russell  Sister Kenny (1946) Mourning Becomes Electra (...
3   Shirley MacLaine[nb 1]  The Apartment (C/M, 1960) Irma la Douce (C/M, ...
4             Tom Hanks  Big (C/M, 1988) Philadelphia (D, 1993) Forrest...
..                  ...                                                ...
71          Edmond O'Brien                                                NaN
72          Lynn Redgrave                                Georgy Girl (C/M, 1966)
73        Omar Sharif[nb 9]                              Doctor Zhivago (D, 1965)
74          Hilary Swank  Boys Don't Cry (D, 1999) Million Dollar Baby (...
75            Jane Wyman      Johnny Belinda (1948) The Blue Veil (D, 1951)

                                  Supporting Role  Total awards  \
0   Kramer vs. Kramer (1979) Adaptation. (2002)              7
1                 Terms of Endearment (1983)              6
2                                         NaN              5
3                                         NaN              4
```



```
jupyter Film Awards Last Checkpoint: 3 minutes ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run Code

In [7]:
# Reading the HTML table into a list of dataframe objects
df = pd.read_html(str(awards_table))
df

Out[7]: [                                                 Film  Year  Awards  \
0                                            Parasite  2019       4
1                                      Ford v Ferrari  2019       2
2     Learning to Skateboard in a Warzone (If You're...  2019       1
3                                The Neighbors' Window  2019       1
4                                        Little Women  2019       1
...                                               ...   ...     ...
1311                                The Yankee Doodle Mouse  1943       1
1312                                       The Yearling  1946       2
1313  Yesterday, Today and Tomorrow (Ieri, oggi, dom...  1964       1
1314                          You Can't Take It with You  1938       2
1315                       Zorba the Greek (Alexis Zorbas)  1964       3

      Nominations
0               6
1               4
2               1
3               1
4               6
...           ...
1311            1
1312            7
1313            1
1314            7
1315            7
```

4.Data Processing:

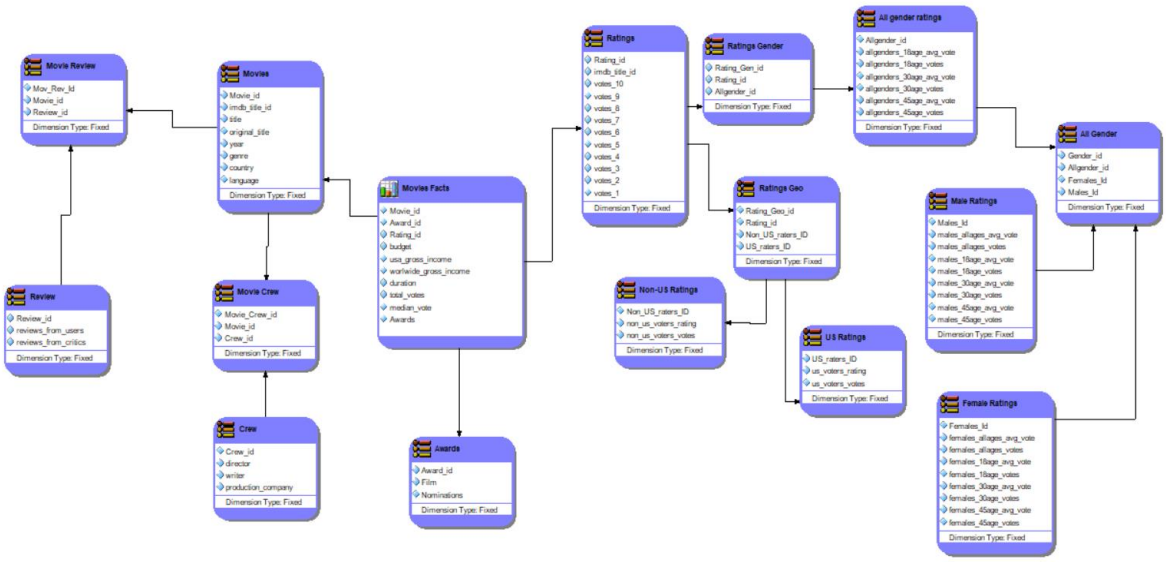
The steps to be followed in the data processing are as follows:



- ER Studio has been used to model the data and build a dimensional model.
- For the ELT process SSIS packages have been utilized. Through this the data has been first loaded into the staging area in Microsoft SQL server management studio.
- From the staging area, data has been loaded in the Data Warehouse.
- Errors were handled using SSIS package, where all the foul data is logged into error outputs with the help of red arrows which will redirect the errors to required destinations which can be SQL tables or flat files.
- With the help of an analysis services project, we built an OLAP cube.
- Data wrangling has been done using Microsoft SQL server management studio.
- For the purpose of visualisation, Tableau is used.

5. Snowflake Schema

The following figure represents the Snowflake schema:

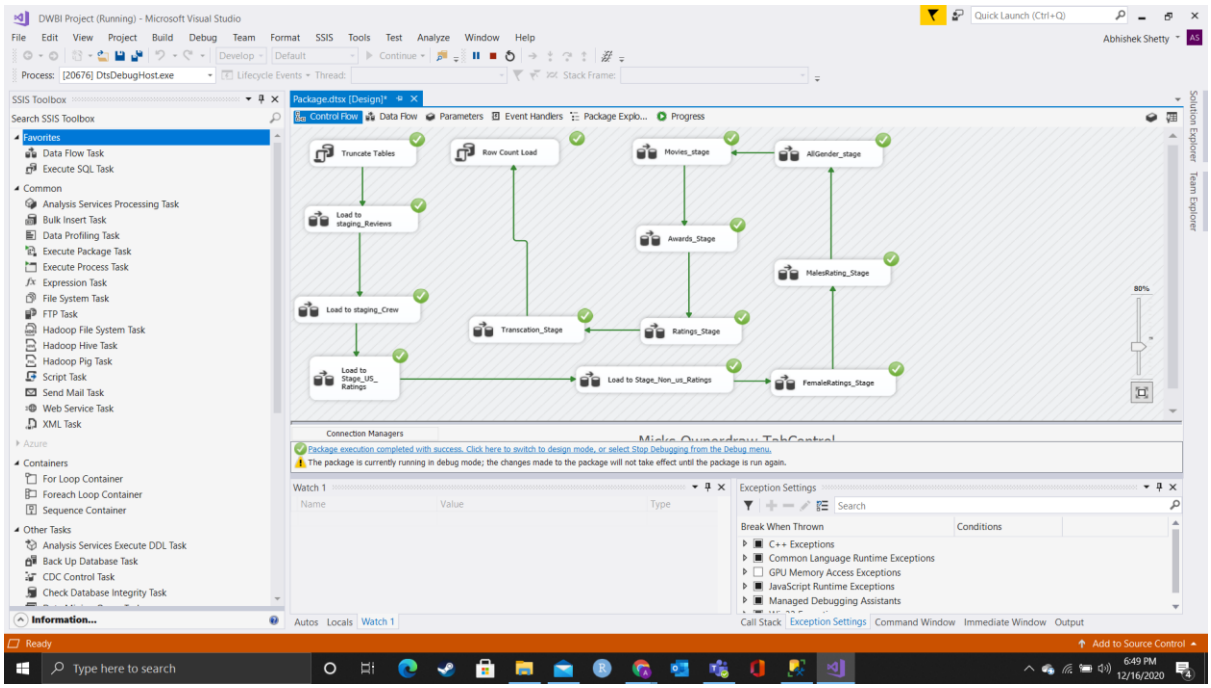


6. Implementation of the Multidimensional Schema

All the tables have been created in the staging area using SSMS by passing appropriate DDLs

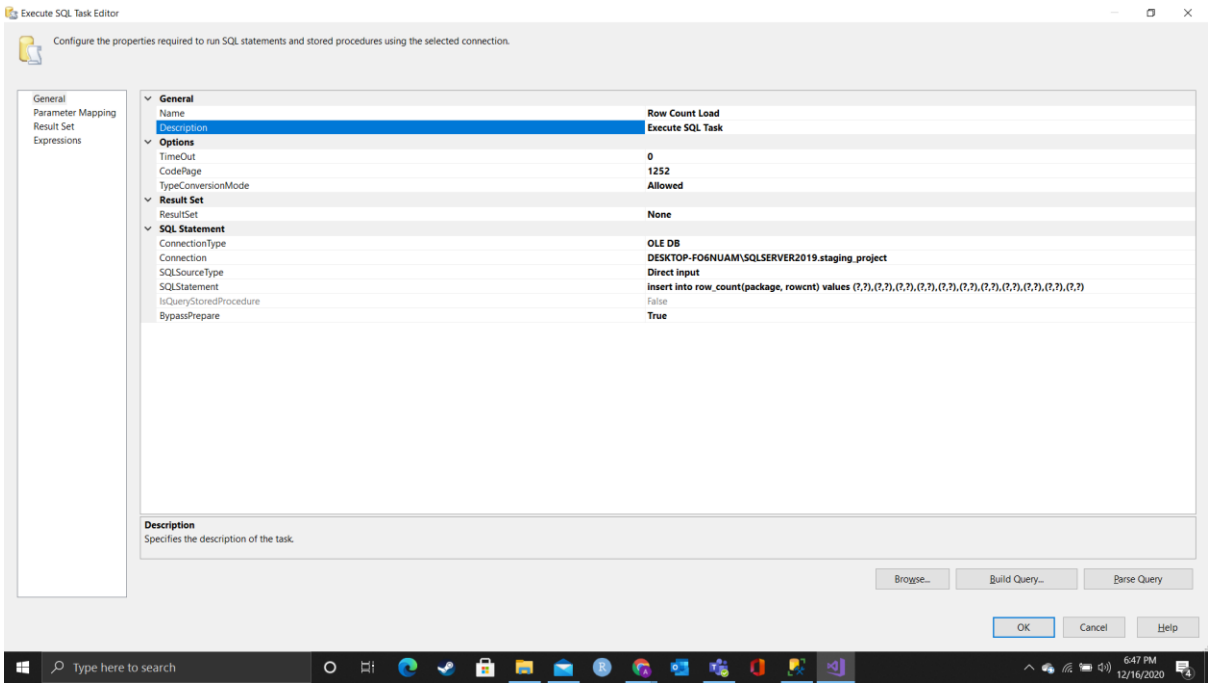
6.1. Loading into Staging Tables:

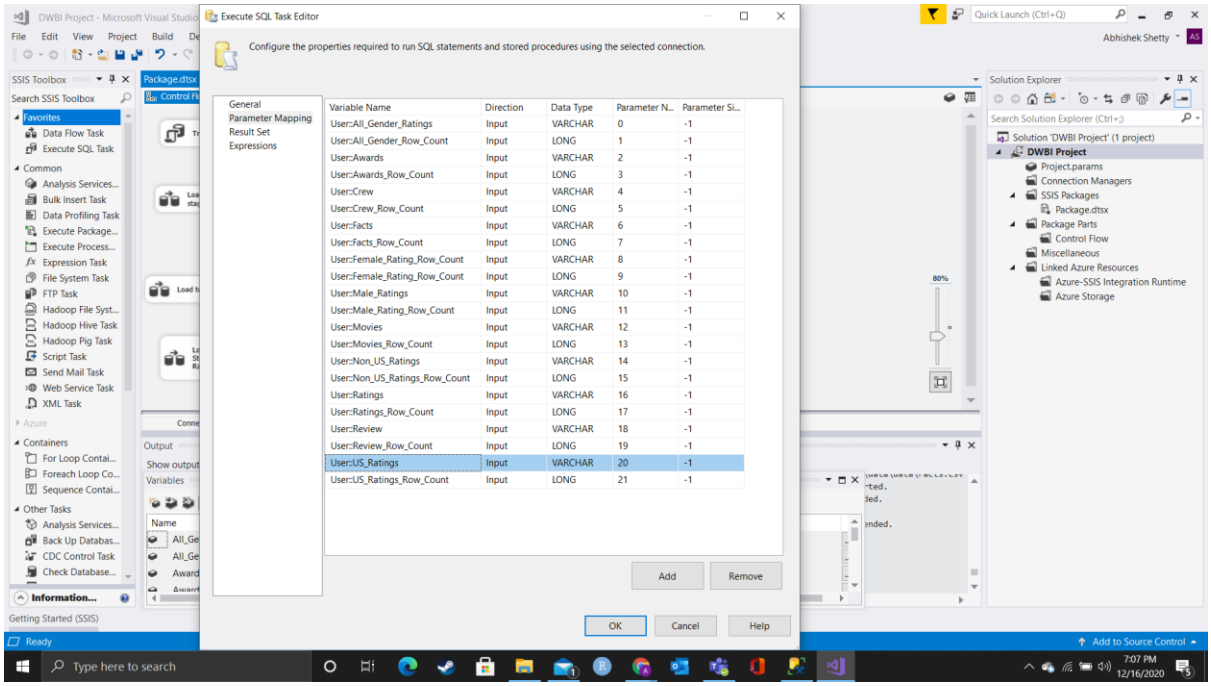
We created a database called ‘staging’ and wrote the DDLs for all the staging tables. The data is then loaded into these tables using SSIS as shown below.

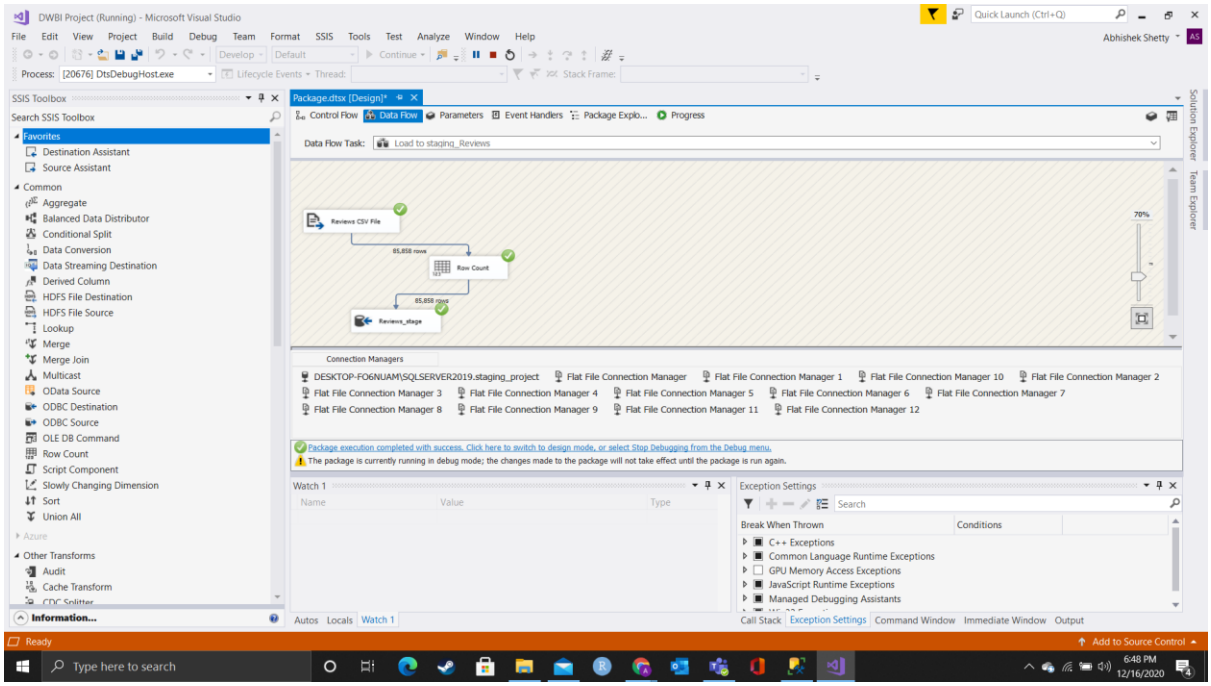


6.1a Tracking Row count of data Sources:

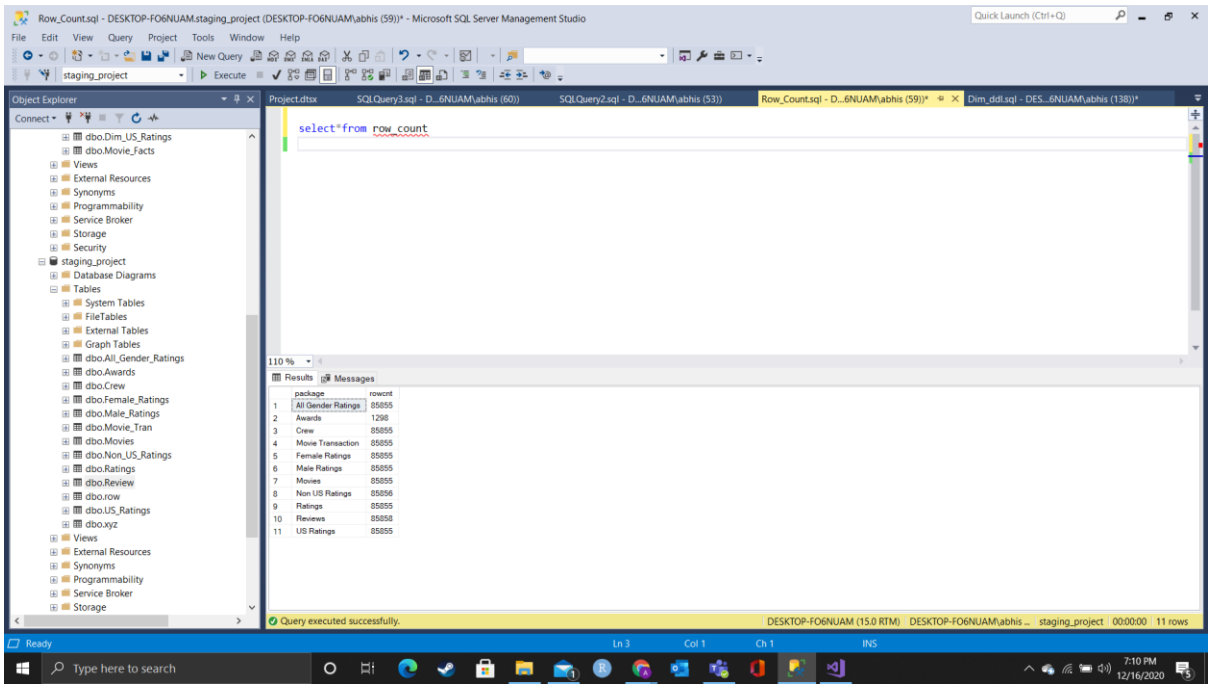
We have used the row count function in SSIS to track the count of the rows present in the flat file. The row count for each of the flat files is stored in a variable we created. Then finally the row count for each of these flat files is then loaded into the table Row_Count, using the Execute SQL Task.







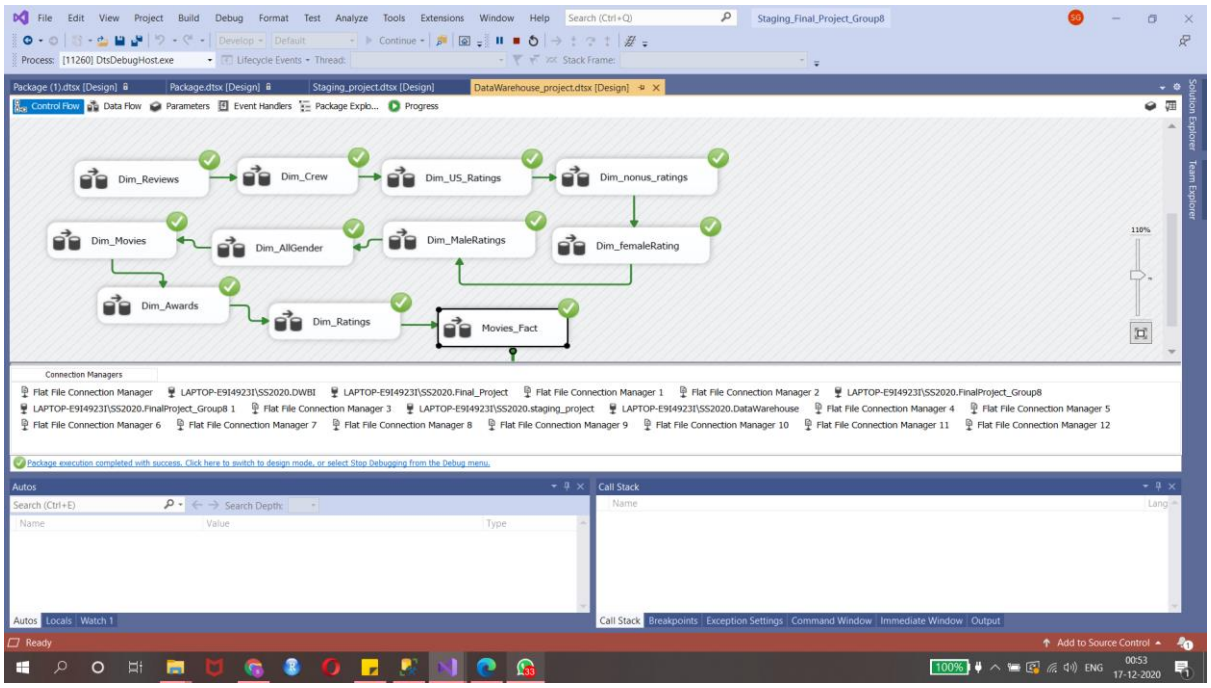
So as shown in the figure below, all the rows have been loaded into the tables from csv files.



6.2 Loading the Data into the Data Warehouse:

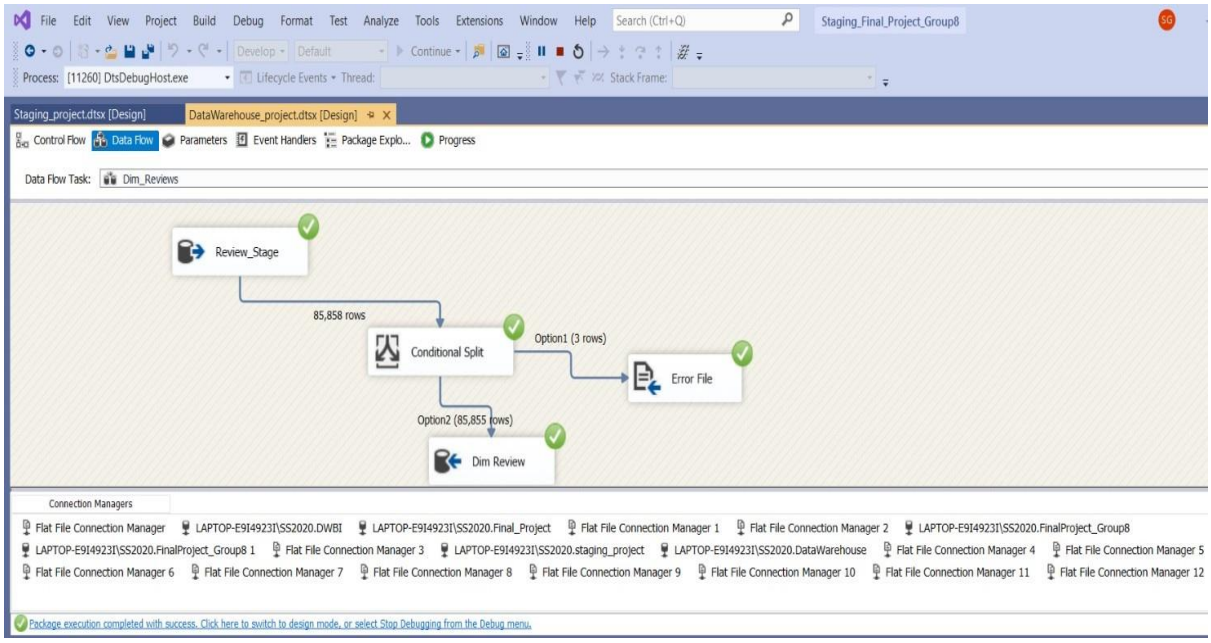
Data Warehouse design plays a very important part in performing BI functions and reporting. The data warehouse has been designed by combining all the data sources and validating the data in order to maintain accuracy.

To load the data into the data warehouse, surrogate keys have been created for all the tables. From the staging area, data has been loaded with the help of an SSIS package which encompassed aspects like error handling, SCD, and lookups in order to get the desired values in the dimensional model



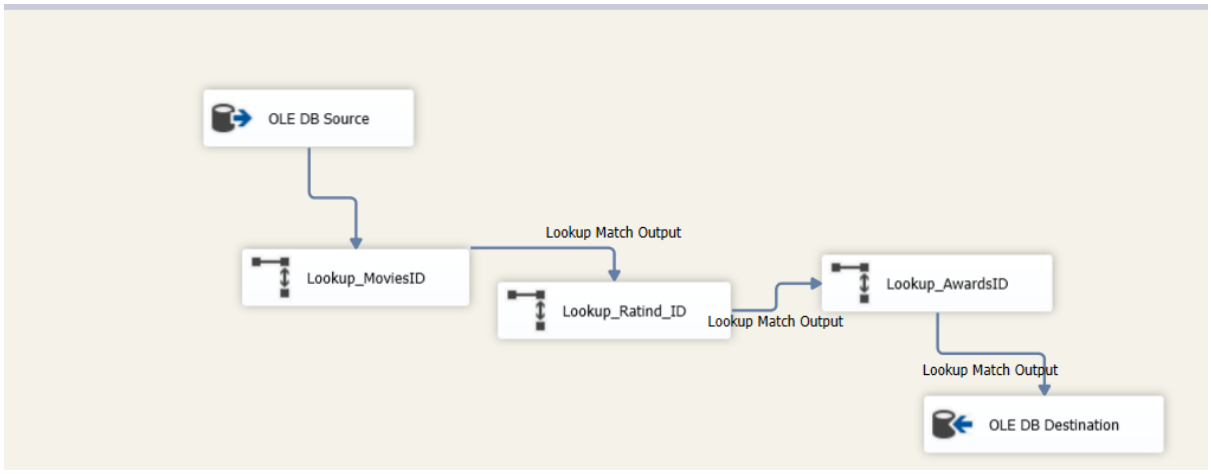
6.2a Error handling:

Filthy data needs to be removed to build a robust data warehouse, therefore, erroneous values have been removed from the data while loading the data from the staging area to the warehouse. The following image gives an example where Review data is being checked for errors i.e if the Review ID is null than the data is removed and moved to the error file.



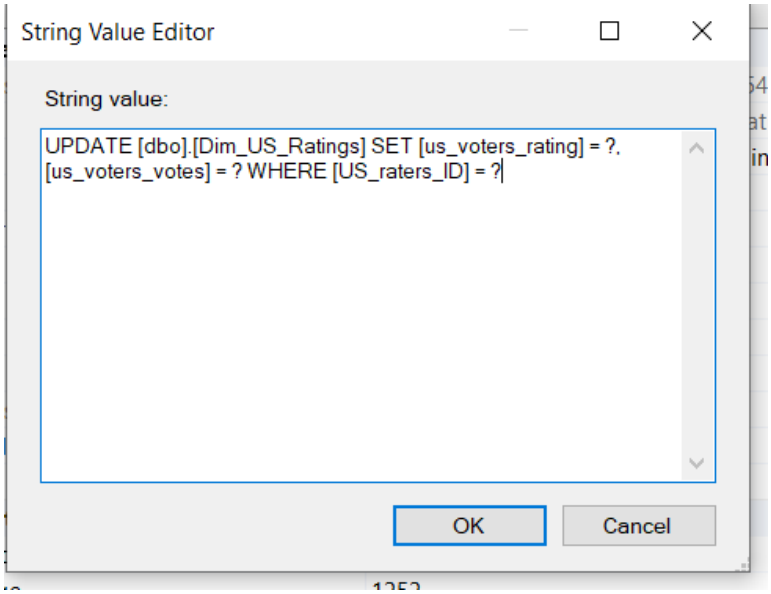
6.2b Lookups

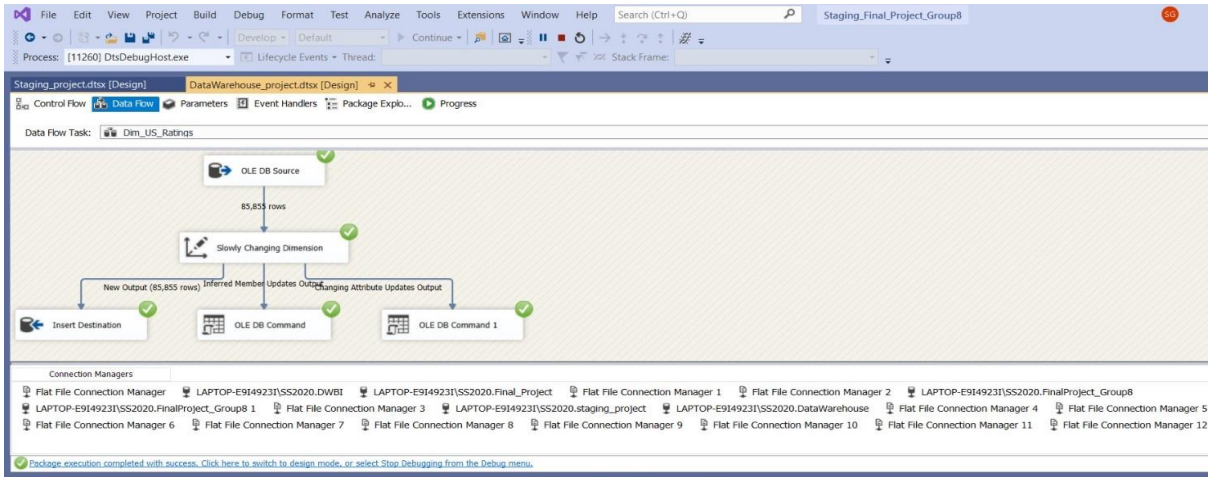
To maintain integrity of the data, lookups have been used in some data flow tasks



6.2c Slowly Changing Dimensions:

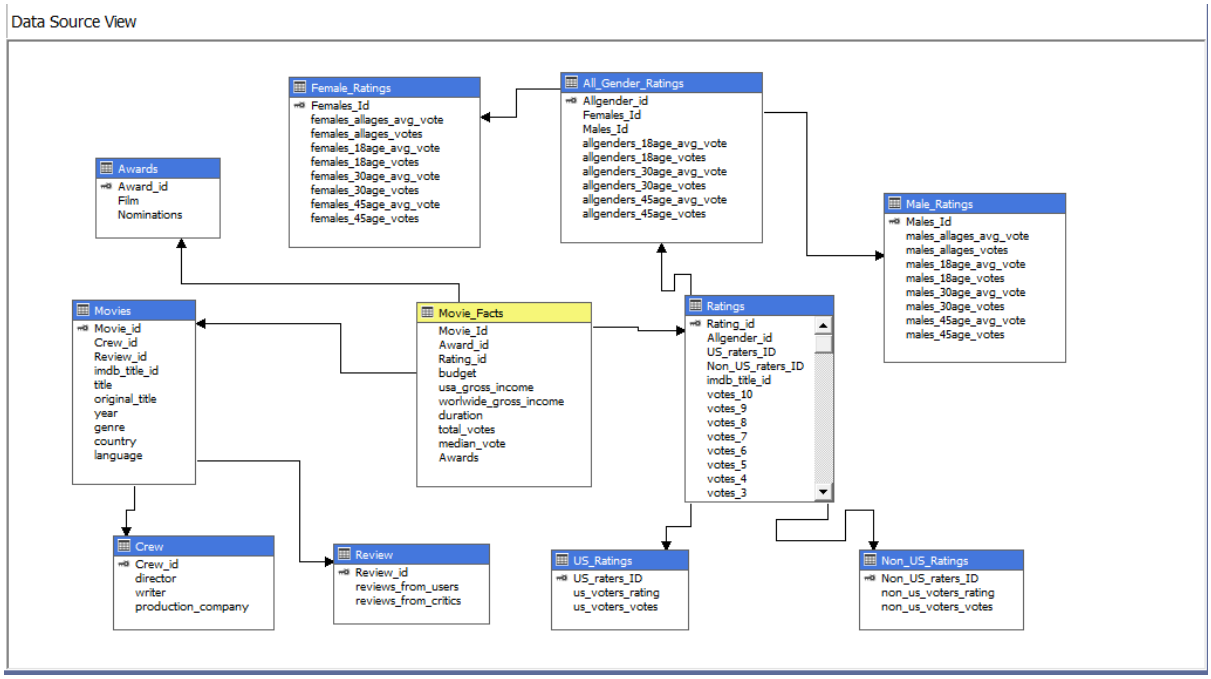
Dimensions tend to change over time and hence we implemented SCD in order to track the data changes that occur. For example, the ratings data keeps on getting updated so to the table Us_Ratings. We have given the following update statement so that it gets updated when new ratings are added to the table.





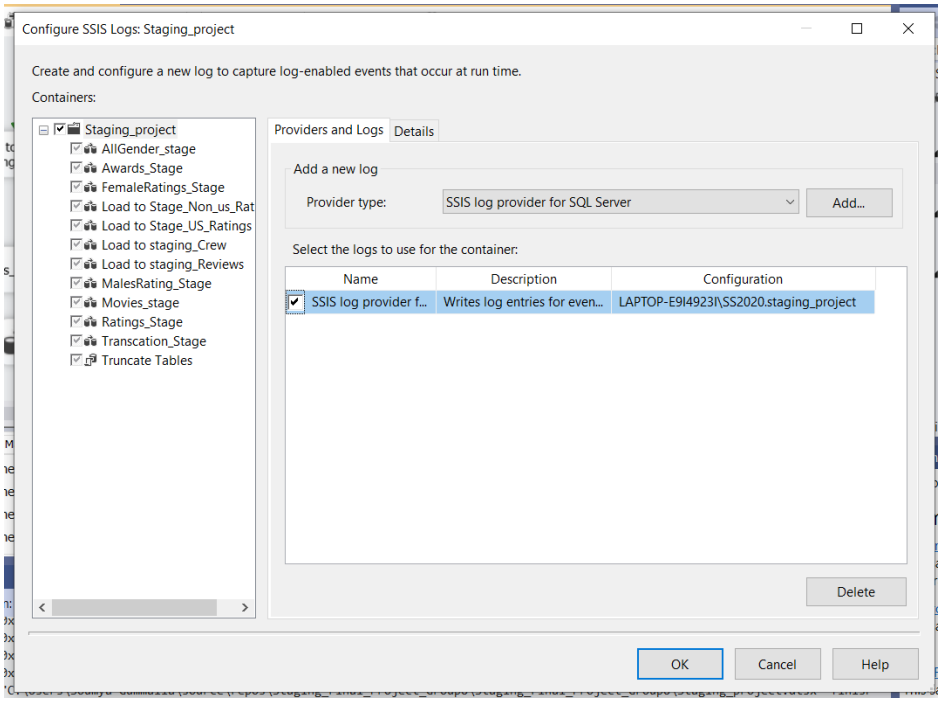
6.3 OLAP Cube:

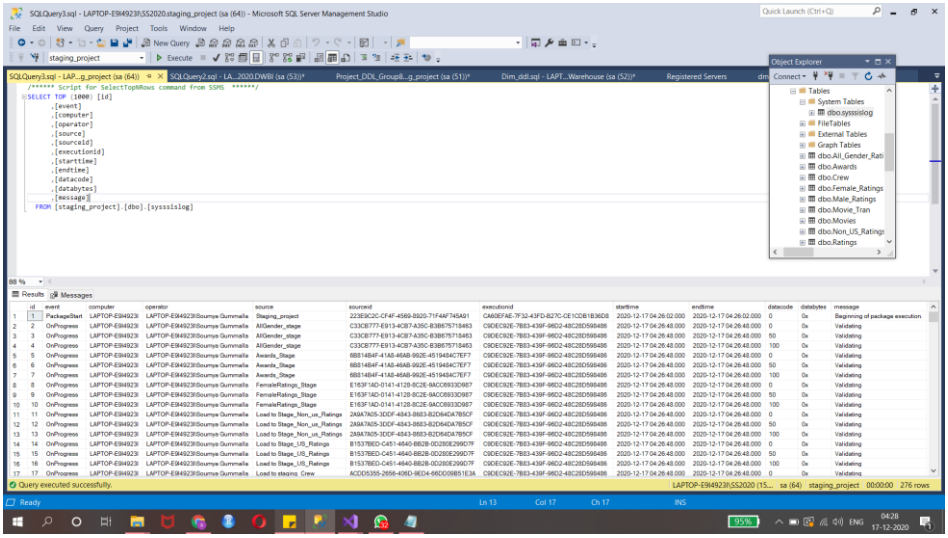
To get a better insight of the data, we processed an OLAP cube using Visual Studio. We then obtained a multidimensional model containing the following tables.



6.4 Logging

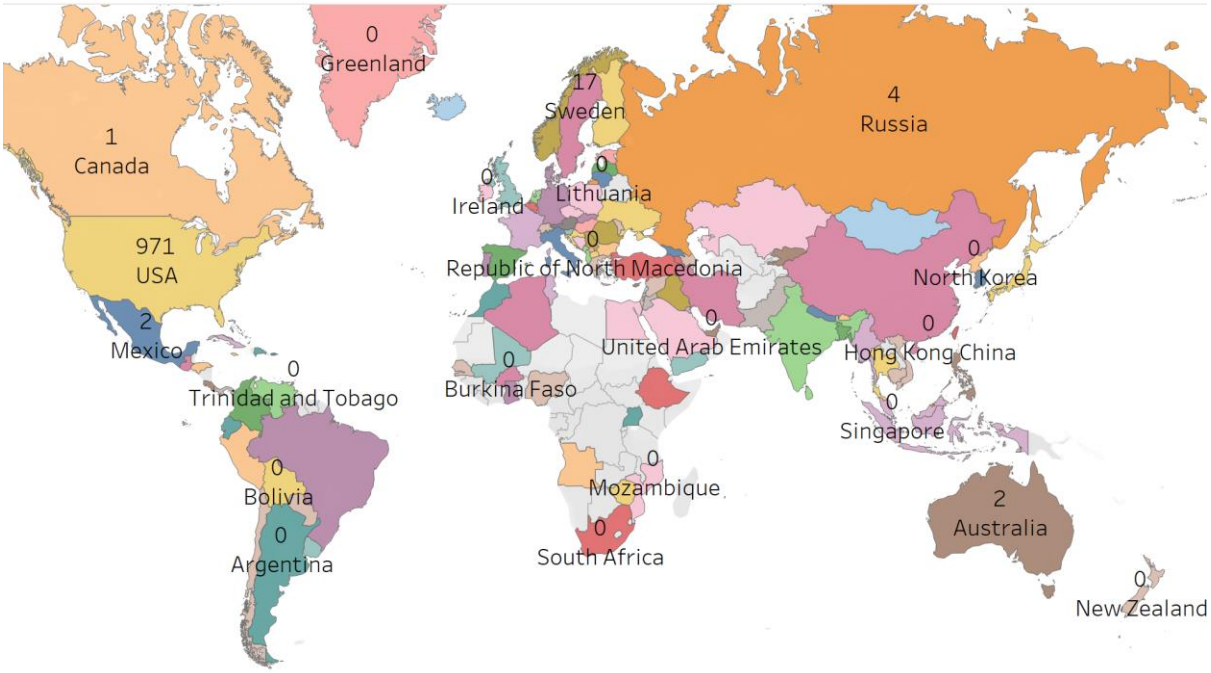
To keep a track of all the tasks that are being executed logging has been performed with help of SSIS and a System Log table has been created as follows:



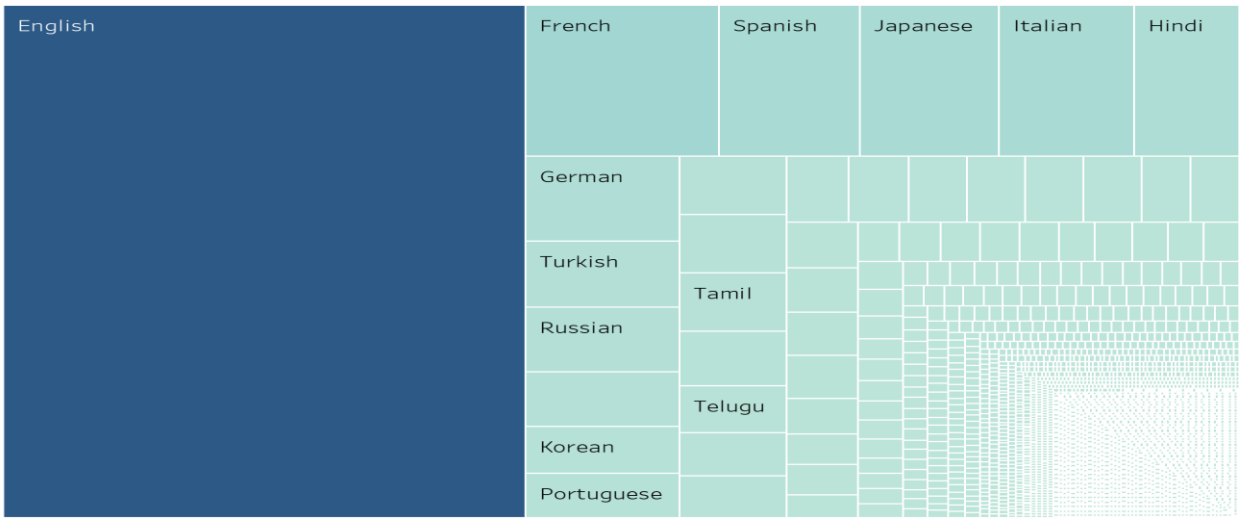


7. Visualizations and Analysis

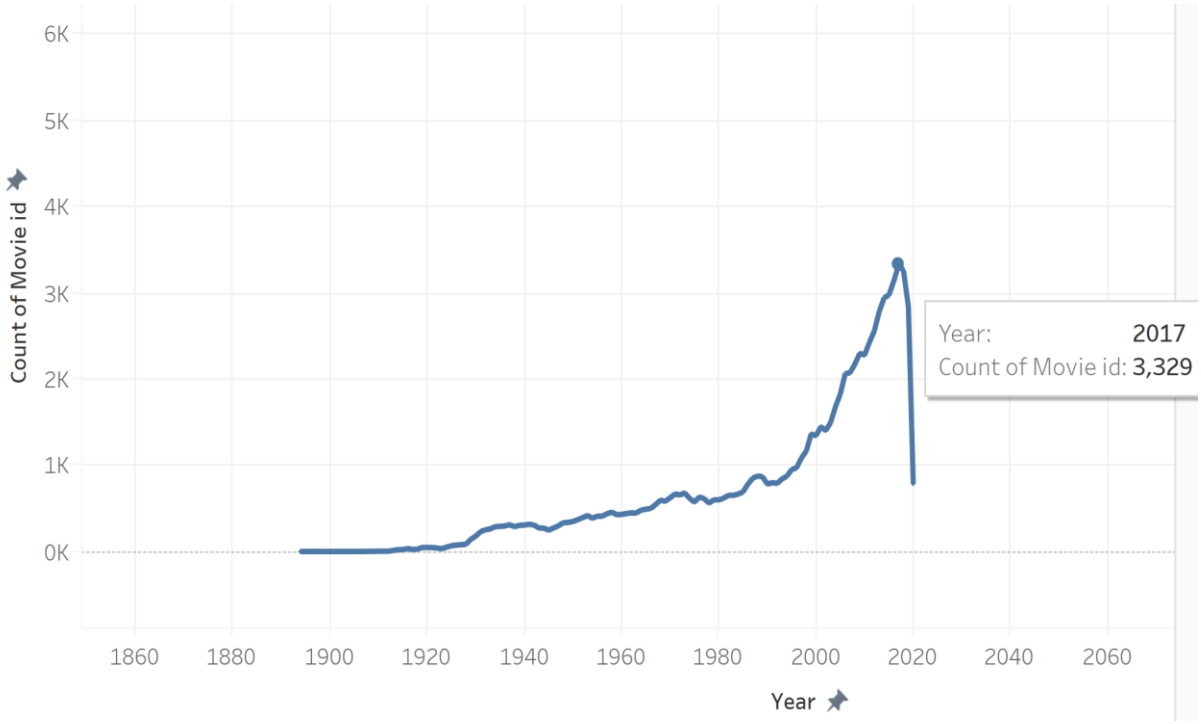
- On visualizing the countries and number of awards won, it was found that the USA stands highest with 971 awards.



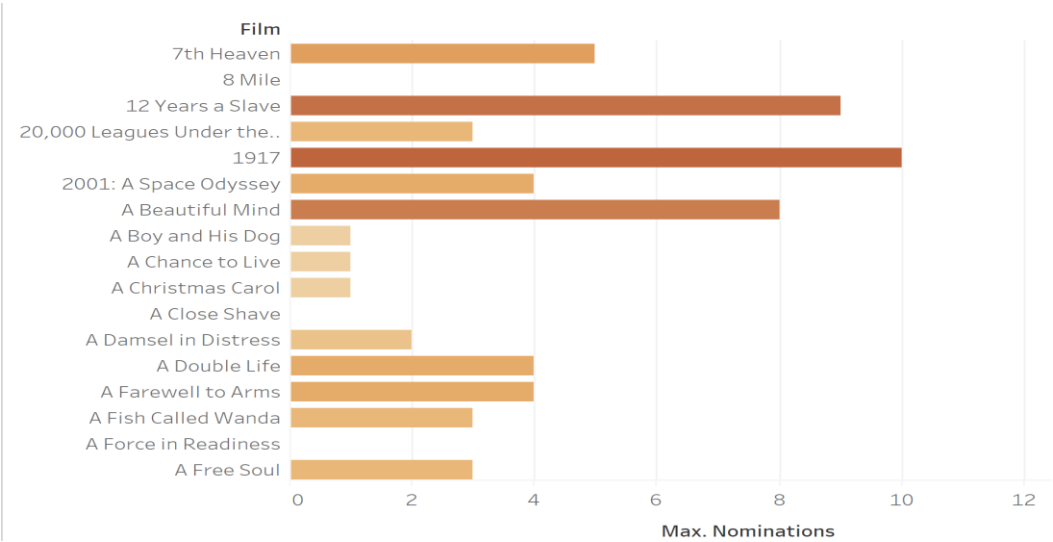
- English has the highest number of ratings compared to other languages.Then comes movies in French,Spanish,Japanese ,Italian and Hindi.



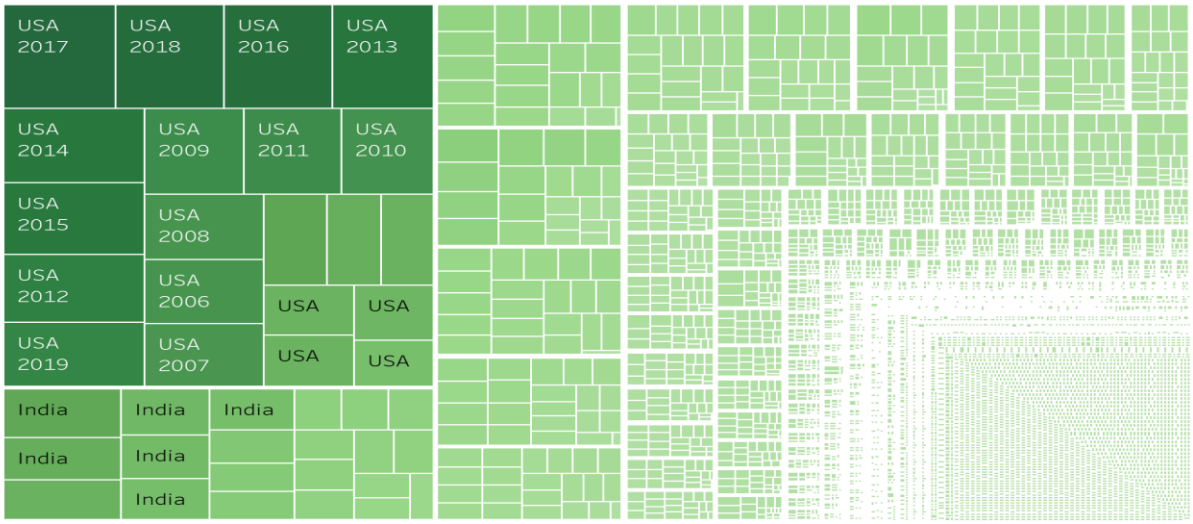
- The number of movies released started increasing after 1990 till the year 2017 when 3329 movies were released. But, after this the number started to decrease as shown in the figure below.



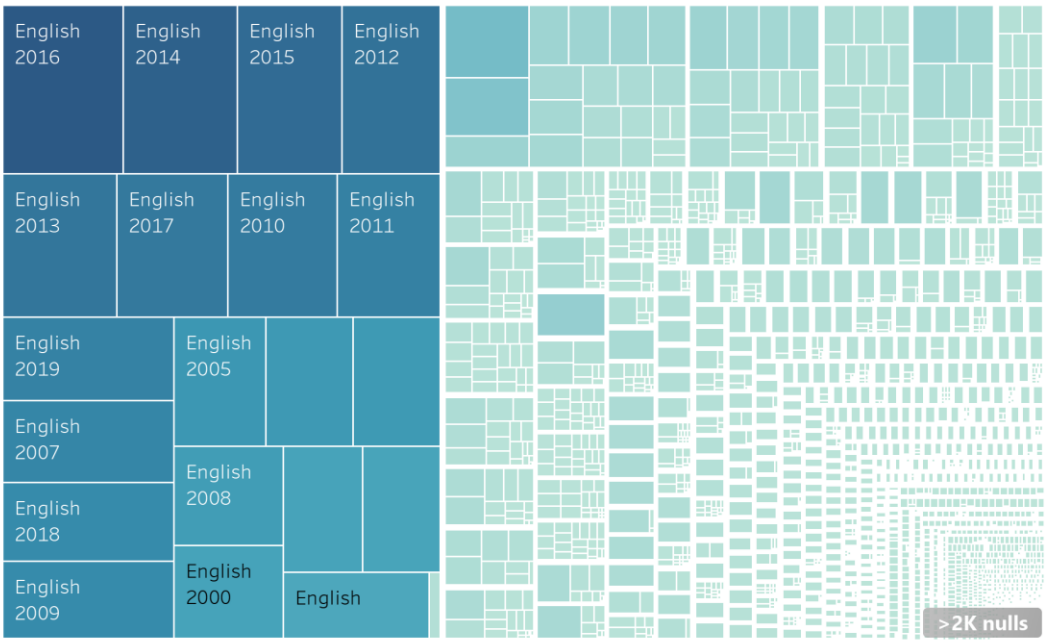
- We analysed film vs nominations data in Awards dataset to see which film got the most number of nominations. From the plot,it can be inferred that '1917' got more i.e. 10 while films like 'A boy and his dog','A chance to live', 'A christmas carol' got least number of nominations for awards.



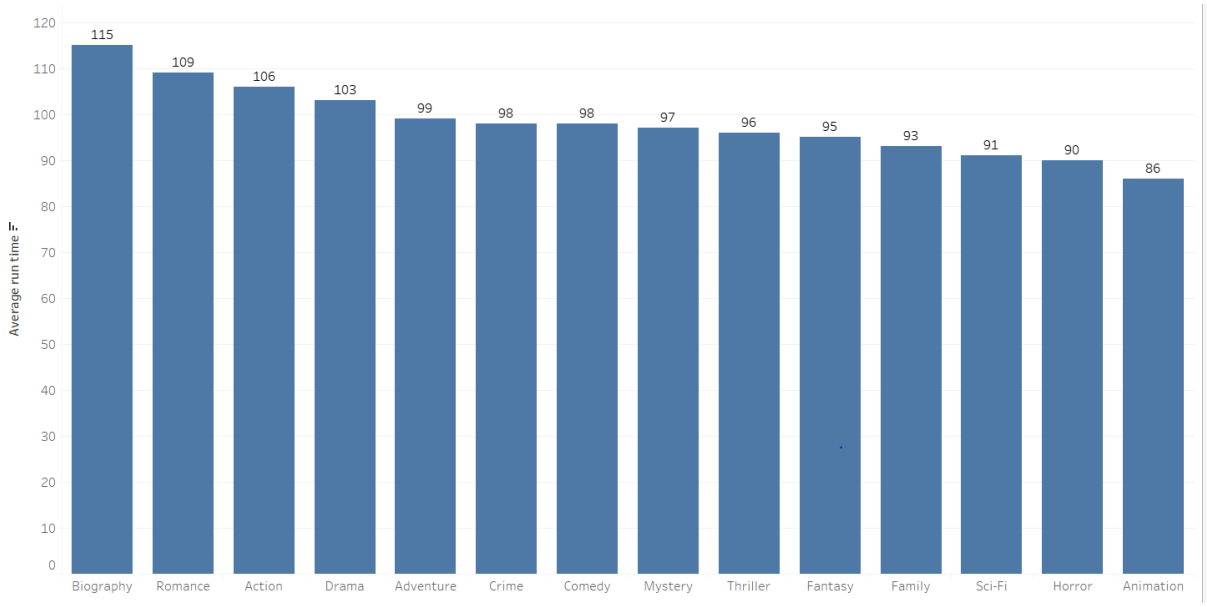
- On visualising the number of movies released from the year 2000 to 2020 in all countries,it is seen that USA stands first and India stands second.Other countries such as France,Georgia,Australia did not release much movies every year.In the figure shown below, each block represents a country and each tile in the block represents number of movies released from the year 2000 till 2020.



- Compared to all other languages, it is observed that English and Spanish movies are watched by many people across the world and hence it is more profitable if movies are released in these languages.



- From the visualization we can infer that Biography movies have the highest average runtime followed by romance and action movies and animation movies have least runtime compared to the movies in other genres.



8. Conclusion

From our analysis, it can be inferred that the following factors contribute to the success of the movie:

- **Language** plays an important role because in this case,most of the films shot in specific languages are viewed more and some languages are not understood by many people.English being an international language is understood by many people across the world and is a good option for directors who want to shoot a film and make profits.
- **Country** in which a film is released also plays a critical role.From our analysis it is found that the USA won many awards and also the number of films released in this country is more compared to any other country. So if a film is released in the USA it is more likely that many people watch and there is a high possibility of it getting nominated for awards.
- **Genre and Average Runtime** is also important because ratings depend on these factors and this is something which influences the viewer's choice.

Data Source:

Movies & Ratings Data

https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

Awards Data

Data has been scrapped from Wikipedia

1.https://en.wikipedia.org/wiki/Golden_Globe_Awards
2.https://en.wikipedia.org/wiki/List_of_Academy_Award-winning_films

Appendix

Professor's Comments after Project Proposal 1

Include a revision history
Not sure you have enough interesting data this is a basic set of data that has minimal interest
That isn't two data marts
You don't describe the data sets
Can you get additional data such as awards etc…
I think you will need more

Professor's Comments after Project Proposal 2

Put my comments at the end of the document
You took my comments about expanding the data section too literally
Error handling section needs work
You shouldn't have my questions in the document…see data warehouse design
Your comments on logging are not sufficient you need to log errors and log the process you can't count on sql server
You need to tell me how you will load the data and how the process will work

Professor's Comments after Design Document 1

Validations should be automated
SSMS is not a visualization tool
Show how you scraped data
You need to describe how you build the start schema and the dimensions