Goal

Our goal was to gather data, clean the data and put the data in SQL. In order to accomplish this goal, we first have to extract the data from Wikipedia and Kaggle. Then we had to clean, parse, sorted and merged the data. I cleaned the movie by removing outliers, dropping duplicates and parsing the data. After cleaning the data sets, I merged the Kaggle csv with the Wikipedia Json. Then I placed the combined data into a SQL database.

Movies-ETL Assumptions

One problem that might come out is when new movies come out. New movies come out on a monthly or even weekly basis. What happens to the data when new movies come out? The data stays the same. Therefore, occasionally someone needs to add the new movies to the data base. But they can’t just add it. They still have to extract it, clean it, parse it and finally add it to the data base.

Another problem/assumption I have for our data base is that, is there enough information. Although we did get our information from two large sites, there may be other sites that have more movies that we do not know about.