**D A T A P R O F I L I N G**

**SQL Staging tables:-**

**Name\_Basics:**

* This table comprises a list of people in the movie industry- actors, directors, writers, producers, etc.
* The attribute ‘nconst’ serves as a primary key, holding a unique value for every individual. The table has birth & death year, and ‘known for titles’ which is a list of movie titles the individual is renowned for.
* There are 8.12% of null values in Primary Profession.
* Birth Year & Death Year have 1.2M & 1.35M null records respectively.

**Title\_Basics:**

* This table has all movies along with details like their genre, runtime, startYear, etc.
* The attribute ‘tconst’ serves as a primary key, holding a unique value for every movie title.
* StartYear has 79k null records, whereas EndYear has all records null.
* 9,302 movies are recorded as Adults, ie. 1.5% of all movies.
* Drama is the most common genre with 18.5% of all titles falling in this category.
* 37% of all titles do not have a recorded runtime.

**Title\_Crew:**

* This table is a list of directors & writers in every single movie.
* Once again, ‘tconst’ serves as the primary key, as every movie has exactly one record.
* The attributes directors & writers do not have actual names, but the respective ‘nsconst’ values.

**Title\_Ratings:**

* This table has ratings and the number of votes for movies.
* All movies have not been listed in this table, as there are only 277k records.

**Title\_Principal:**

* This table lists the crew of every movie (actors, directors, producers, cinematographer, etc.). It also records the name of characters played by the actors. However, the attribute ‘character’ has 65.6% null records.
* ‘Job’ has 82.8% null records.
* The profession with highest number is actors (1.17M)

**Title\_Akas:**

* This table lists out every movie in every region. This explains the high number of records.
* Out of 2.6M records, 531k have region value null.
* The columns ‘language’, ‘types’ & ‘attributes’ also have high null records.

**TSV Files:-**

All TSV files have the same schema, where the primary key is a combination of ‘tconst’ & ‘date’. These files show the box office collection of the movies for every single day they were screened in theatres.

**Avatar tsv file:** In this particular TSV file Rank column has the highest number of null values with 176 null values represented as dash(-)

**The Avengers:** In this particular TSV file Rank column has the highest number of null values with 25 null values represented as dash(-) and in %LW and %YD with null percentage of 15.46%, 4.12%

**The Avengers - Age of Ultron:** In this particular TSV file %LW and %YD with null percentage of 10%, 2%

**The Avengers - Infinity War:** In this particular TSV file Rank column has the highest number of null values with 61 null values represented as dash(-) and in %LW and %YD with null percentage of 5.67%, 1.42%

**The Avengers - End Game:** In this particular TSV file Rank column has the highest number of null values with 35 null values represented as dash(-) and in %LW and %YD with null percentage of 5.67%, 1.42%

**Black Panther:** In this particular TSV file Rank column has the highest number of null values with 70 null values represented as dash(-) and in %LW and %YD with null percentage of 4.55%, 1.44%

**Spider Man:** In this particular TSV file Rank column has the highest number of null values with 37 null values represented as dash(-) and in %LW and %YD with null percentage of 6.25%, 1.56%

**Star Wars:** In this particular TSV file %LW and %YD with null percentage of 6.67%, 1.67%

**Titanic :** In this particular TSV file Rank column has the highest number of null values with 54 null values represented as dash(-) and in %LW and %YD with null percentage of 9.43%, 1.51%