DATA CONSISTENCY, CLEANING, AND QUALITY STRATEGIES

1. We are using tReplace component to replace "\N" from source tsv files (which indicates null) with null to result in SQL null in all the required tables. We are also using tReplace to identify and remove other characters like "$", "\"", "-", "," etc. from the number, decimal, string, and currency fields in all the required tables as part of data cleaning.

   Used in these tables:- stg_imdb_brands_gross, stg_imdb_franchises_gross etc.

2. We are using tUniqueRow to extract only the unique rows from the source tsv files based on the column values selected in the component and remove duplicates.

   Used in these tables:- stg_imdb_name_basics, dim_imdb_titleType etc.

3. We are using tFilterRow to extract only the required rows based on the condition so that only the applicable rows will be used in the integration job to improve job performance and reduce execution time.

   Used in these tables:- dim_imdb_genres, dim_imdb_genres_ml_rejects etc.

4. We are using tConvertType to convert the mismatching column datatypes from the data sources to database tables. tReplace component requires the input column datatypes to be of string datatype, so we are using tConvertType to convert other datatypes to be replaced to string, making the changes in the tReplace component, and converting the datatypes back again.

   Used in these tables:- stg_imdb_name_basics,

5. In stg_movies_box_office_worldwide, we identified that the movie title in the top 1000 movies tsv file is "Lemony Snicket's A Series of Unfortunate Events", whereas in the IMDB dataset, the movie's primaryTitle is "A Series of Unfortunate Events". We have handled this by making the data from source tsv consistent to the database.

6. In stg_imdb_franchise_list, for the franchise Walden, its top ranking movie was present in the tsv file as 'Chronicles of Narnia: The Lion, the Witch, and the Wardrobe', which was not matching with the title in the IMDB dataset. We corrected the title using tReplace.

7. We are handling the repeating groups in the integration stage using the tNormalise component in tables like dim_imdb_title_basics_genres, dim_imdb_title_crew_writers, dim_imdb_title_crew_writers etc. for removing repeating groups.