REJECT STRATEGIES IMPLEMENTED

In Staging:

1. Stg_box_office_worldwide

In this table, we are considering the first level of accepted top 1000 movies records to be the ones with the matched column values of primaryTitle and startYear from stg_imdb_title_basics to Title and Year column values from the top 1000 movies tsv file respectively; unmatched ones are rejected in the first level. Then, the second level of accepted top 1000 movies records is obtained by mapping the originalTitle and startYear from stg_imdb_title_basics to Title and Year values from the top 1000 movies tsv file respectively; unmatched ones are rejected in the second level. Then, the third level of accepted top 1000 movies records is obtained by mapping the startYear from stg_imdb_title_basics to the column Year incremented by 1 from the top 1000 movies tsv file to correct the data in Year and startYear columns in both sources; unmatched ones are rejected in the third level. Then, the fourth level of accepted top 1000 movies records is obtained by mapping the startYear from stg_imdb_title_basics to the column Year decremented by 1 from the top 1000 movies tsv file to correct the data in Year and startYear columns in both sources; unmatched ones are rejected in the fourth level and this is the final level of rejects.

In Integration:

1. Dim_imdb_genres_ml_rejects

In this table, we are considering the rejected records to be the ones present in MovieLens dataset and not present in IMDB dataset. We are achieving this by mapping the genres column from stg_ml_movies to genres column in dim_imdb_genres and performing a join on movieId column from dim_imdb_title_basics (lookup) with movieId from stg_ml_movies (main) and catching the rejected values by setting the “Catch lookup inner join reject” to true in the dim_imdb_genres_ml_rejects table.

2. Dim_imdb_name_basics_knownForTitles_rejects

In this table, we are considering the rejected records to be the normalized knownForTitles column values in stg_imdb_name_basics_knownForTitles not matching with the tconst column values in dim_imdb_title_basics. We are achieving this by mapping the nconst column from stg_imdb_name_basics_knownForTitles to nconst column in Dim_imdb_name_basics and performing an inner join on knownForTitles column from dim_imdb_name_basics_knownForTitles with tconst from dim_imdb_title_basics and catching the rejected values by setting the “Catch lookup inner join reject” to true in the m_imdb_name_basics_knownForTitles_rejects table.

3. Dim_imdb_title_akas_rejects and Dim_imdb_title_principals_rejects

In these tables, we are obtaining the rejected records by setting the “Catch output reject” as true in the tMap of corresponding reject tables.

4. Dim_imdb_title_crew_directors_rejects and Dim_imdb_title_crew_writers_rejects

In these tables, we are considering the rejected records to be the ones with the unmatched tconst and nconst values from dim_imdb_title_basics and dim_imdb_name_basics respectively with the tconst and nconst values from stg_imdb_ title_crew_directors/ stg_imdb_ title_crew_writers tables.

5. All fct_ml_rejects tables

In these tables, we are considering the rejected records to be the ones with the unmatched movieId column values from the corresponding stg_ml tables with the movieId column values from dim_imdb_title_basics; an inner join is performed and the “Catch lookup inner join reject” value is set to true in the all the fct_ml_rejects tables.