Evolving Graphs for IMDB using Portal

Aim of project:

Graphs are used to represent a plethora of phenomena, including the Web, social networks, biological pathways, and semantic knowledge bases. The phenomena that are represented by these graphs can change over time. Researchers progress in this area depends on the availability of standardized datasets on which performance can be tested.

Scope of Work

In this work, we make progress towards studying IMDB database as a temporal graphs to represent evolution of graph topology and of vertex and edge attributes.

Survey on Previous Work

Bipartite graphs, random graphs model, property graphs were some of the graphs were used to represent the the IMDB data set. However, these graphs were inadequate to describe the social processes occurring within the dataset and were not scalable and efficient for graph queries.

Research Findings

We used IMDB dataset from imdb website dating from 1900 -2020, which includes around 6 million movies and 5 million people associated with them.

1.Graph types:

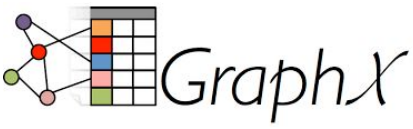
We generated graphs to define relationship with actors and movies such that:

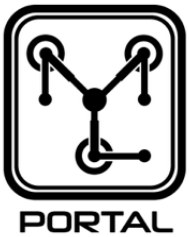
a. **Movie-person**
   Movies and Person as nodes of graphs and there is an edge between them if they are work together.
   Edge validity - from years they appear first till 2020

b. **Collaboration graph**
   here we consider nodes of graphs as person and edge between them if they have collaborated for a movie.
   Edge validity - from years they appear first till 2020

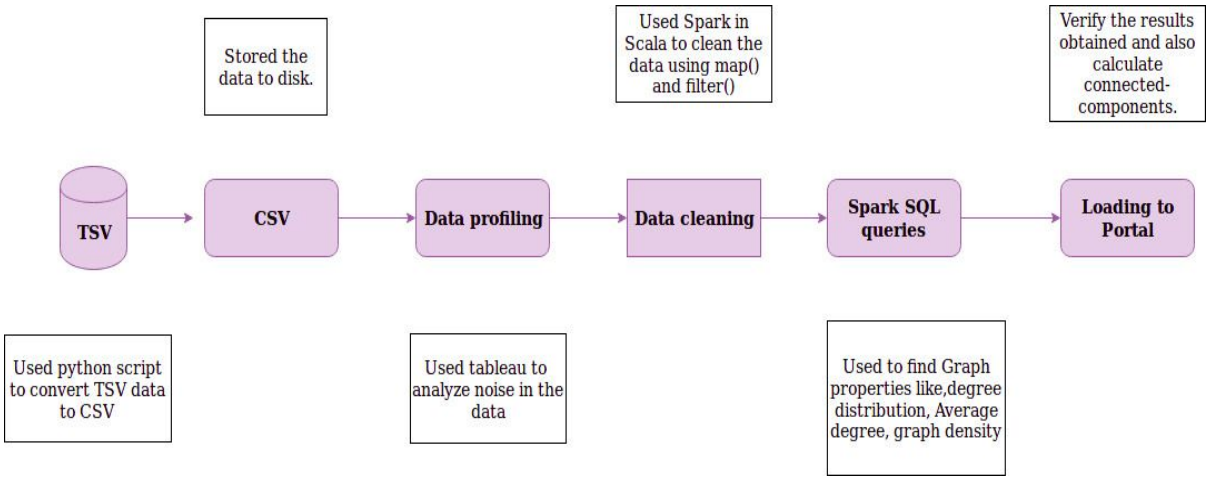
Edge Validity-This is an temporal graphs where structure of graph is changing w.r.t time.

2. Software used:



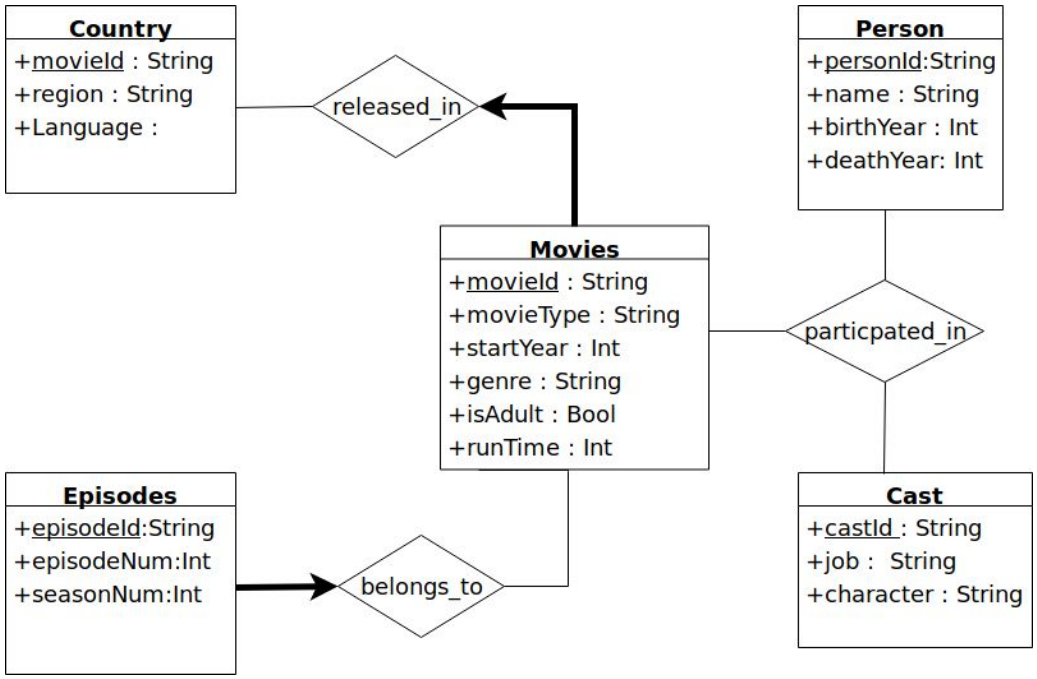


4. Analysis Methodology:

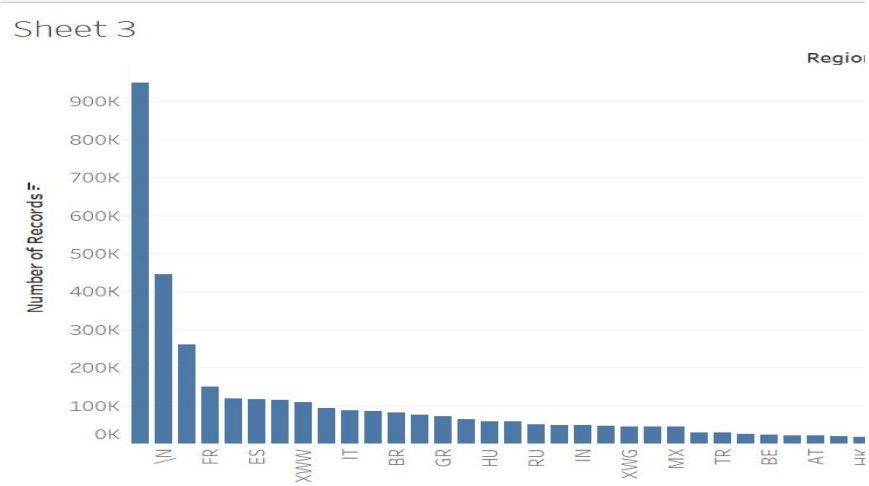


a. *Data Profiling :*

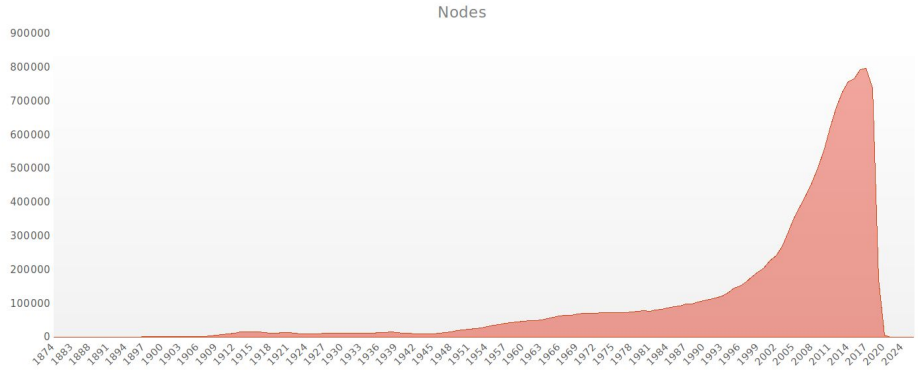
It important to understand structure between different entities if our data, so ER Diagram was used.



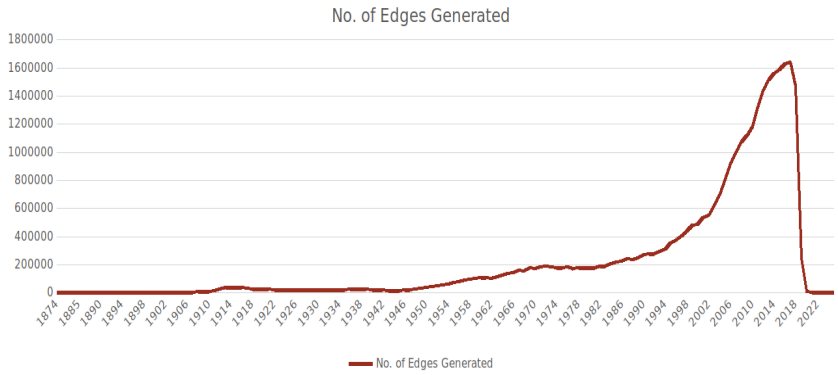
Used tableau to get eagle view of the data, and evaluate the type of each attributes.



Geographical analysis of the movie where most of the movie belong to US while second largest movies are “null” values



Node generation for person and movie over time

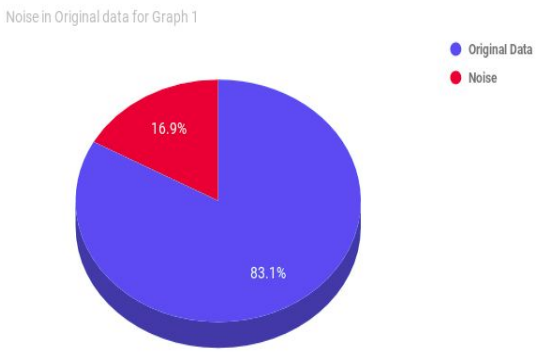


Edges generation over time

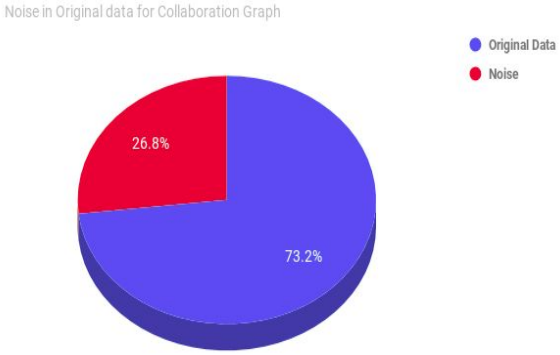
b. data cleaning :

Used Scala's map, filter and reduce functions to clean the data.
Uncleaned data contains:

1. Null values
2. Start Year is missing
3. Semantic noise like Int value in Column with schema of type string.



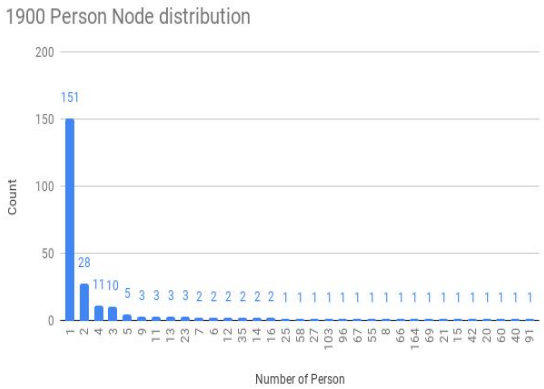
Distribution for data cleaning method for movie-person graph



Distribution for data cleaning method for collaboration graph

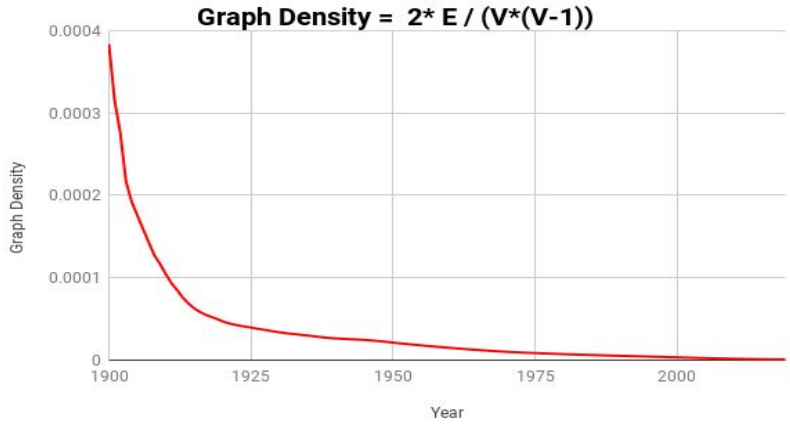
c. Results from Spark Queries.

Here, we have used Spark Queries to find the graphs statistics like node generation over time, edge generation over time, degree distribution..



It used to to evaluate the edges from person node , from above we can observe that most of the people in 1900 were involved in just one movie while one person was maximum involved with

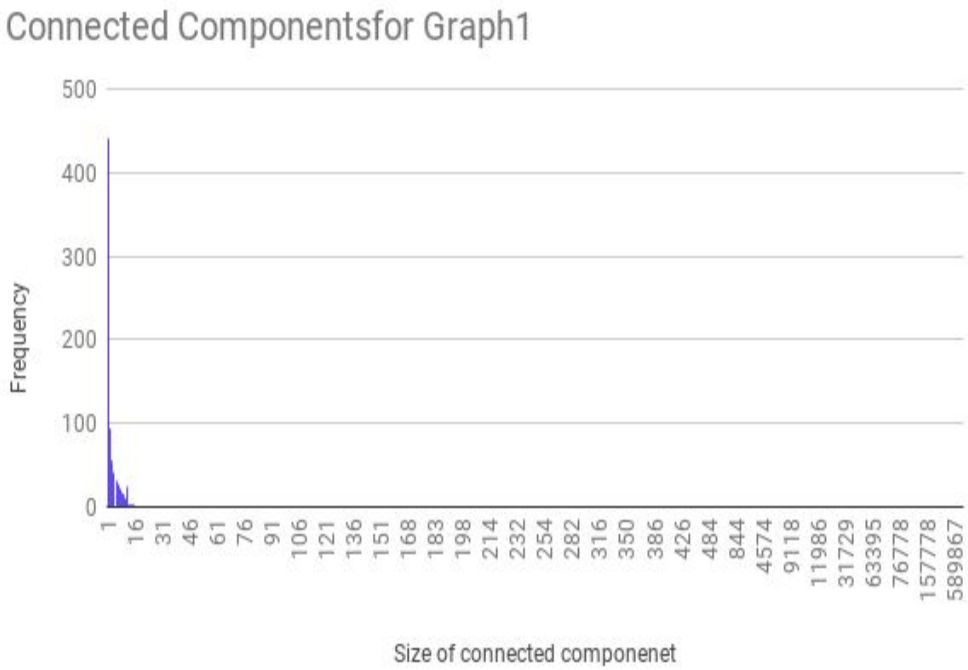
164 movies. While in 2015 maximum people were involved with only one movies, hence this trend is followed.



Above diagram shows that graph density in decreasing over time, which suggest that graph is becoming more sparse.

d. Loading data to PORTAL:

Finally, data was loaded to Portal, to verify our results and find interesting Statistics like connected components, Page rank, triangle count.



Connected components follow power law which is similar to social media graph.

Conclusion:

By observing the properties of graphs from portal and Spark queries we can say that IMDB database closely follows Online Social network graph. Queries from portal can give us information about the social processes in IMDB dataset.