CSC 501 ALGORITHMS AND DATA MODELS

University of Victoria, 201909

ASSIGNMENT 01

Work by:

Trishala Bhasin, Nirav Galani, Karan Tongay

1. **INTRODUCTION**

In this document, we present our findings from and the process of working with the data provided as part of the assignment using relational databases. We have used the following tools

* Python3
* Pandas
* Tableau

We started with exploring the data using the small dataset (100,000 ratings) and created a relational database using this small dataset. Subsequently we used the 20M dataset to seek out our final insights and to demonstrate scalability. Details of our decision making and final outcomes and visualizations are discussed in subsequent sections.

The following files are also associate with this report

* data\_csv\_preprocessing.ipynb
* model.ipynb
* revised\_tags.csv
* revised\_ratings.csv
* genres.csv
* links.csv
* revised\_movies.csv

1. **DATA MODELING**

When reviewing the original data, it was evident to us that some pre-processing was necessary. After considering various ER designs, we determined that the optimal ER diagram would be the one shown in Figure 1 of the appendix. We considered various factors, such as potential maximum table sizes and consequential joining complexity and normalization.

For example, let us consider ER diagram in figure 2 vs figure 1 using the small dataset. We know that there are 610 users, 3683 tag applications, 9742 movies, 100836 ratings and 20 genres. Considering this data, the complexity of the tables in the 2 ER diagrams would be as follows:

The pre-processing was done using the file data\_csv\_preprocessing.ipynb. This results in 5 files which would represent the data to be used in our database.

* revised\_tags.csv
* revised\_ratings.csv
* genres.csv
* links.csv
* revised\_movies.csv

This data was then used to create the data model using the file model.csv. Querying for the data and the use visualization was done using Tableau.

Both the pre-processing and the modeling were done first on the small dataset and then, to demonstrate scalability, on the 20M dataset. We discuss the scalability and associated algorithmic aspects in section 4.0

1. **TRANSFORMING THE RAW DATA INTO INSIGHTS**
2. **ALGORITHMIC DESIGN**
3. **RELATIONSHIP TO MODERN PRATICE**
4. **CONCLUSION**
5. **APPENDIX 1 - Figures**

TAGS

Associated with

have

have

1

1

1

LINKS

RATINGS

MOVIES

Associated with

GENRE

Figure 1: Final ER diagram