CSCI 620/Sections 01-03/Rivero, Introduction to Big Data, Spring 2205

# Assignment 1 – Relational databases

## Description

Let's model IMDB (Internet Movie Database). You can find a description of the dataset here: https://www.imdb.com/interfaces/. Use the following files available in myCourses:

- name.basics.tsv.gz
- title.basics.tsv.gz
- title.crew.tsv.gz
- title.principals.tsv.gz
- title.ratings.tsv.gz

The database system we are going to use is MySQL Community Server version 5.X.

## Your tasks (100 points)

Create the following relations, attributes and keys (attributes in bold denote primary keys):

- Movie(**id**, title, isAdult, year, runtime, rating, votes)
- Genre(**id**, name)
- MovieGenre(**mid**, **gid**)
  - mid FK to Movie(id)
  - gid FK to Genre(id)
- Person(**id**, name, birthYear, deathYear)
- Actor/Director/Producer/Writer(**pid**, **mid**)
  - pid FK to Person(id)
  - mid FK to Movie(id)

You need to explore the given files in order to understand their contents. Since these files are large, you should create a Java program to explore them. You can find an example on how to manipulate GZip files in the software template. (Hints: We are interested only in movies and TV movies; explore the categories in the 'title.principals.tsv.gz' file; self should be treated as actor; there are additional directors and writers mentioned in the 'title.principals.tsv.gz' file.) You need to provide the appropriate attribute types and load only the expected data. You should not create foreign keys as data loading will take much more time than the expected by the grading software. Instead, you need to guarantee that the inserted data is coherent and valid by performing queries at the end of the process. There are additional memory and time requirements. Check the grading software for more info.

## Submission instructions

- Use the software template provided in myCourses.
- Submit a single ZIP file to myCourses that must be named as your RIT user, e.g., crrvcs.zip. Do not include '@rit.edu.' The file must contain a folder named 'IMDBToSQL' containing your Gradle project.
- Everything will be graded on a Linux machine, so you must always use the exact names provided in this document, software template and grading software.

## Grading rubric

- Check the grading software.