Assignment 1 – Hadoop

Louella Creemers

641347

Github Link: https://github.com/LouellaCreemers/HadoopAssignments

# Requirements

* Python
* MrJob installed

# How to run

Add the u.data file from the movielens dataset to the Hadoop VM. Add the sum\_rating\_per\_movie.py file to the same folder.

In this folder where both these files are located, run the following command:

*python sum\_rating\_per\_movie.py u.data* for exercise 1

*python sum\_rating\_per\_movie\_sorted.py u.data* for exercise 2

# 6 – Sum ratings given for each movie

**Functions:**

A screenshot of a computer

Description automatically generated with medium confidence

*mapper\_get\_ratings* create the needed key/value pairs of a movieId and 1. The number 1 is the amount of rating given per row. We are going to sum up all those 1’s in the next step.

*Combiner\_count\_ratings* sum up all the values of ratings with the same keys (movieID).

*reducer\_count\_ratings* sum up all the keys which are left after the combiner.

**Final result:**

The final result is a list of unsorted rows with a unique movie ID and the amount of ratings that belongs to it

**A picture containing table

Description automatically generated**

# 8 – Sort movies by amount of ratings

**Functions:**

**A screenshot of a computer

Description automatically generated with medium confidence**

*mapper\_get\_ratings* is the same as the last exercise

*combiner\_count\_ratings* is the same as the last exercise

*reducer\_count\_ratings* We need a key value of countedratings/movieId now instead of movieid/ sum of counts because it needs to be sorted. The current movieID isn’t helpful anymore, which is why we return None.

Reducer\_sort\_counts is going to actually sort the counts by creating a loop. This will make the movie with the most ratings show first. The int(movieID) is in there to keep the result readable.

**Result:**

A list of amount of counted ratings with a movieid next to it. The movie with id 50 has the most ratings with an amount of 583

**A picture containing table

Description automatically generated**