# Large Scale Data Management

Map-Reduce
Lab Guide 1

2020/2021

Consider the IMDb dataset available at `https://www.imdb.com/interfaces/`.

**Steps**

1. Compute the number of films for each genre from `title.basics.tsv.gz`.
2. Use a combiner to lessen the amount of data to the reducer.
3. Measure the time it takes to do the computations for different prefixes of the files.
4. Compute a list showing the rating for each title using also `title.ratings.tsv.gz`.
5. Use chained jobs to produce a list of films with rating of at least 9.0, sorted by rating.

**Questions**

1. How do solutions with MapReduce compare with those from Lab Guide 0?
2. What is the impact of combiner and sorting stages?

**Learning Outcomes** Formulate Hadoop Map-Reduce jobs for simple aggregation queries. Recognize the overhead of setting up and deploying distributed solutions.