Middle East Technical University

Department of Computer Engineering

CENG 495

Cloud Computing
Spring 2022–2023
HW - 3

Due date: 2023-06-06 23:59

1 Introduction

In this homework, you will use Apache Hadoop's MapReduce to get insights from the Movie Industry dataset. You will use the Java language for this homework.

2 Setup

Setup Hadoop in Pseudo-Distributed operation mode, using the single node cluster approach. You can follow the Hadoop tutorial here and the MapReduce tutorial here.

3 Task

Download the dataset and extract the `.csv` file. This is the only input file you will need for this assignment. You might want to preprocess the dataset to make your job easier for the later tasks (e.g. convert the `csv` to a `tsv`). If your Java program expects a preprocessed dataset, make sure to include a script that takes the original dataset and converts it to the one your program expects.

I recommend using visidata to inspect the dataset.

3.1 Tasks

Report on the following;

- The amount of time (in minutes) it would take to watch every movie in the dataset, back to back (**total**)
- The average runtime of the movies (in minutes) (**average**)
- How many times each actor has been top-billed (starred) in a movie (**employment**)
- Average number of IMDb votes on G, PG, PG-13 and R rated movies (**ratescore**)
- The average IMDb score of genres that have more than 9 movies (**genrescore**)

4 Submission

- Use Java programming language using the Apache Hadoop library.
- Archive your project as a `.tar.gz` file and name it as "firstname_lastname.tar.gz".
- This is an individual assignment. You can discuss your ideas with your peers but using implementation specific code that is not your own is strictly forbidden and constitutes as cheating. This includes but not limited to friends, any previous homework, CENG homework repositories on GitHub, or the Internet in general. The violators will get no grade from this assignment and will be punished according to the department regulations.

Your code will be evaluated using the Pseudo-Distributed local mode of Hadoop. Your submission will be extracted, and the following commands will be executed on the top level of your submission:

```
# compilation
hadoop com.sun.tools.javac.Main *.java
jar cf Hw3.jar *.class

# running
hadoop jar Hw3.jar Hw3 total <input.csv> output_total
hadoop jar Hw3.jar Hw3 average <input.csv> output_average
hadoop jar Hw3.jar Hw3 employment <input.csv> output_employment
hadoop jar Hw3.jar Hw3 ratescore <input.csv> output_ratescore
hadoop jar Hw3.jar Hw3 genrescore <input.csv> output_genrescore
```

If you want to deviate from the commands given above within reason, drop a `README` file explaining how to build & run your project and I will use that during grading.