Middle East Technical University

Department of Computer Engineering

CENG 495

Cloud Computing

Spring 2021–2022

HW - 3

Due date: 2022-06-18 23:59

1 Introduction

In this homework, you will use Apache Hadoop's MapReduce to get insights from the Top Hits from Spotify 2000-2019 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. I can also recommend ArchLinux Wiki's Hadoop entry, but specific instructions will differ for your distribution.

3 Task

Download the dataset. 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 (i.e. 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 will not be using datasets in the submission archives. I recommend using visidata to inspect the dataset.

3.1 Tasks

- List the total duration (ms) of each song in the dataset (**total**)
- List the average duration (ms) of each song (**average**)
- List how many times an artist has been in the top hits list (**popular**)
- Separate the songs according to whether they are *explicit or not*, then list the average popularity for the songs in those two separate categories. Output explicit songs in `part-r-00000` and non-explicit songs in `part-r-00001` (**explicitlypopular**)
- Partition the songs by year; first partition is the songs that came out on or before 2002 (year $\leq$ 2002), the second partition is for songs that came out between 2002 and 2012 (2002 $<$ year $\leq$ 2012) and final partition is the songs that came out later than 2012 (2012 $<$ year). Report the average danceability of these 3 partitions (part-r-00000 to part-r-00002). (**dancebyyear**).

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 popular <input.csv> output_popular
hadoop jar Hw3.jar Hw3 explicitlypopular <input.csv> output_explicitlypopular
hadoop jar Hw3.jar Hw3 dancebyyear <input.csv> output_dancebyyear
```

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.