An Overview of the Problem and the Data sets

Problem statement

Write and execute a MapReduce program to figure out the top 100 trending songs from Saavn’s stream data, on a daily basis, for the week December 25-31. Although this is a real-time streaming problem, you may use all the data till the $(n-1)_{th}$ day to calculate your output for the n_{th} day, i.e. you may consider all the stream data till 24 December (included) in your program to find the trending songs for 25 December.

Definition of trending

The term ‘trending songs’ may be defined loosely as those songs that have gathered relatively high numbers of streams within small time windows (e.g. the last four hours) and have also shown positive increases in their stream growth rates.

How is a stream defined at Saavn?

A stream is a record of a user playing a song. Each stream is represented as a tuple with the following attributes:
(song ID, user ID, timestamp, hour, date)

Each tuple consists of the song ID of the streamed song, the user ID of the user who streamed the song, the timestamp (Unix) of the stream, the hour of streaming, and the date of streaming.

Data

- The file https://s3.amazonaws.com/mapreduce-bde/part-00000 contains one month(December) of stream records. Please note that this file is huge (~44GB) and will consume a lot of your internet bandwidth if you choose to download it onto your local machine.
- The file https://s3.amazonaws.com/mapreduce-bde/trending_data_daily.csv contains the trending songs for each day of December, as calculated by Saavn. You may compare your output with these and improve your algorithm to obtain a better match.

- The file https://s3.amazonaws.com/mapreduce-bde/saavn_sample_data.txt contains a sample of 10 million stream records from the original dataset. You may use this to run simple jobs and get an idea of the data.

Further details regarding the problem statement, the method of execution and grading will be added in a few days.