Distributed Data Management – Spring 2024

Lab #2 – Map Reduce

Data:

10k_view_data.csv contains data about ‘viewing events’ - tuning a TV to a station.

This is a small sample of the data you will use in your future work on the course.

- mso_code – unique 5-digit ID number for each MSO (חברת כבלים)
- device_id – unique ID number for each device or set-top-box (ממיר)
- event_date– date the viewing event occurred in the format “YYYYMMDD”
- event_time – time the viewing event occurred in the format “HHMMSS”
- station_num – ID of each TV station
- prog_code – ID of each program broadcasted in each event

Exercise:

For this exercise, we will define “Prime-Time” as the time between 20:00 and 23:00 (excluding 23:00:00).

- Solely from **stations with even ID** codes, find the **most viewed program during Prime-Time** hours.
- Return the **amount of viewers (viewing events)** that that program had, along with its **program code**.

Guidelines:

- Use mrjob library (`!pip install mrjob`).
- You may use Pandas or similar libraries for data-exploration / viewing purposes **only**, all calculations on the data should be done **using MRJob exclusively**.
- You may use additional library as you wish (itertools, collections, re, etc.).
- You are allowed to execute a **single** MRJob (that is, read the data once).
- Assume the input file is in the executing directory.
- The output should be in the format of: <number_of_viewers> <program_code> for example, a sample output should look like: 12 “AB012345678910”

Submission:

- An `.ipynb` file containing your code
- A PDF version of the executed `.ipynb` code
- The `.py` that your `.ipynb` file creates, if you choose to do so (as seen in tutorial)
- Do NOT package the files in a .zip file or any other folder.

Extra Challenge (not for submission):
Think of more complex queries and try executing them on the data only using MRJob.