Distributed Data Management – Spring 2024

Lab #3 – Spark

Data:

Lab3_view_data.csv contains data about 'viewing events' - tuning a TV to a station.

This is a small sample of real data which you'll use in your future work on the course. The data is a small sample of a few days from November of 2015. In this lab we wish to find the top 5 most active devices during Prime-Time.

- 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). You only need to output stage 3, but 1 & 2 are here to help you.

1. Find the number of viewing events for each device in **each day** during Prime-Time.
2. Use the number of viewing events of each device during Prime-Time to calculate the **average amount** of viewing events of each device during Prime-Time **across all dates** in the data.
3. Return the top 5 devices (device_id) and their average amount of viewing events during Prime-Time across all the **dates in the data**, of the devices with the highest averages.

Guidelines:

- Use pyspark library. You are allowed to use **only spark RDD** (you are not allowed to use spark dataframes).
- You may use Pandas or similar libraries for data-exploration / viewing purposes **only**, all calculations on the data should be done **using pyspark exclusively**.
- You may use any additional library you wish (itertools, collections, re, etc.).
- Assume the input file is in the executing directory.
- The output should be in the format of: <device_id> <average_viewing_events>
  for example, a sample output should be 5 rows that look like: "000001b2c3d" 1.3

Submission:

- An `.ipynb` file containing your code
- A PDF version of the executed `.ipynb` code
- An HTML version of the executed `.ipynb` code
- Do **NOT** package the files in a .zip file or any other folder.