The City College of New York Department of Computer Science

THE GRADUATE CENTER CITY UNIVERSITY OF NEW YORK

DSE I2450/CSc 84030: Big Data Analytics/Scalable Computation

SPRING 2021

Homework 2 – MapReduce with MRJob

Problem Statement: we will be solving the same problem in HW 1, the Consumer Complaints challenge from InsightDataScience, but using MapReduce and the MRJob package. Please refer to the original problem statement on GitHub for additional information.
https://github.com/InsightDataScience/consumer_complaints

INPUT:
Your code will be evaluated against the original data set (in CSV format) downloaded from:
https://www.consumerfinance.gov/data-research/consumer-complaints/#download-the-data

The file is roughly 1GB and **unsorted**. The header is still included in the file. A smaller version of the file is also available on Blackboard. You can use the sample file for testing your code.

NOTE: this CSV file contains multiple-line records, so it is expected if your code cannot account for several records lying on the boundary of blocks.

OUTPUT:
Since the output of a MapReduce job is key/value pairs. Please leave the value empty and set the key to the actual output record separated in commas. In other words, if we combine all input together, we should expect a valid CSV file. The output CSV data does not have to contain a header line. In this homework, you are **required to output the records in the sorted order** as requested by the challenge.

SUBMISISON:
The final hand-in should be a single Python file, named **BDM_HW2_LastName.py** that must be executed through command line.