The City College of New York | Department of Computer Science

CSC 445: Big Data Management and Analysis

FALL 2020

Homework 3 – MapReduce

Problem Statement: we are greatly inspired by the Consumer Complaints challenge from the popular InsightDataScience. In fact, we are going to tackle the same challenge but using MapReduce. Please read through the challenge at (the most important sections for us are "Input dataset" and "Expected output"): https://github.com/InsightDataScience/consumer_complaints

Requirements:

1. You must perform your computations using only Python and the MRJob package that we use in class. No external packages, e.g. *pandas*, are allowed.
2. Your code must be able to run as a stand-alone MRJob application.

INPUT:

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

The original data set is roughly 1GB but the sample file is only 4MB, and is available on our class resources under **Data Sets> complaints_sample.csv**. You can use this file for testing your code within a notebook if you prefer.

NOTE: this CSV file contains multiple-line records. Please pay attention to this when reading the data.

OUTPUT:

You are required to write to the standard output in CSV format. Basically, you have to organize each of your record as a CSV row when you output from Spark. The output does not have to contain the header line.

SUBMISISON:

The final hand-in should be a single file, named **BDM_HW3_LastName.py** that takes exactly 1 argument for the input path. Output will be handled through redirection.

```
SAMPLE RUN:
python BDM_HW3_LastName.py complaints_sample.csv > output.csv
```