The City College of New York

Department of Computer Science

CSC 445: Big Data Management and Analysis

FALL 2020

Homework 4 – Apache Spark

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

You are asked to write a program that can read a file of similar format on HDFS, and use Spark to compute the *Expected output* and write them to a folder on HDFS (not to a local file system). Both input and output paths must be specified through command line (similar to MRJob assignment).

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 is available for your access on HDFS at: **/data/share/bdm/complaints.csv** The header is still included in the file. For your convenience, a smaller version of the file is also 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 a folder, where each part must be in CSV format. In other words, if we issue a "getmerge" Hadoop command on the folder, we should expect a valid CSV file. Basically, you have to organize each of your record as a CSV row when you output from Spark. The output CSV data does not have to contain a header line.

SUBMISISON:

The final hand-in should be a single Python file, named **BDM_HW4_LastName.py** that takes exactly 2 arguments for input path and output path, respectively. Your code will be run with 2 executors, 5 cores per executors.

```
SAMPLE RUN:
spark-submit --num-executors 2 --executor-cores 5 BDM_HW3_Vo.py
/tmp/bdm/complaints.csv output_folder
```