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 2020

Homework 2 – MapReduce

Due: 5:30 PM, Mar 11, 2020

Problem Statement: Given a sale data set, e.g. **sale.csv**, similar to the table below:

Customer ID	Transaction ID	Date	Product ID	Item Cost
129482221	T29518	2018/02/28	A	10.99
129482221	T29518	2018/02/28	B	4.99
129482221	T93990	2018/03/15	A	9.99
583910109	T11959	2017/04/13	C	0.99
583910109	T29852	2017/12/25	D	13.99
873803751	T35662	2018/01/01	D	13.99
873803751	T17583	2018/05/08	B	5.99
873803751	T17583	2018/05/08	A	11.99

Note: The data is sorted by the **Customer ID**, and a product could be priced differently across transactions.

Your task is to write a script to produce a CSV file like the following table, **grouped** by **Product ID**:

Product ID	Customer Count	Total Revenue
A	2	32.97
B	2	10.98
C	1	0.99
D	2	27.98

where:
Customer Count = the number of unique customers that bought the product with the given ID
Total Revenue = the total cost of the product in all transactions

Constraints:

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 Python application.

Your submission: The final hand-in should be a single Python file, named *BDM_HW2_<LAST_NAME>py* that takes exactly 2 arguments in the following format:

```
python BDM_Lab2_xyz.py <INPUT_CSV>
```

<INPUT_CSV> is the full path to your input data, e.g. sale.csv. The output will be printed to the standard output where, for example, we could be run as follows:

```
SAMPLE RUN:
python BDM_HW2_xyz.py sale.csv > output
```

Note: the input CSV file will be without header (otherwise, the contents will be the same as the file on our class resources). For testing purposes, you can use sale_small_without_header.csv. The output can be tab separated.