The City College of New York | Department of Computer Science

CSC 445: Big Data Management and Analysis

FALL 2020

Homework 1 – Streaming

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, **sorted** 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 Python only. No external packages, e.g. *pandas*, are allowed.
2. The data set is assumed to be really large. Please do your best not to load everything in memory.

Your submission: The final hand-in should be a single Python file, named *HW1_streaming.py* that takes exactly 2 arguments in the following format:

```
python HW1_streaming.py <INPUT_CSV> <OUTPUT_CSV>
```

<INPUT_CSV> is the full path to your input data, e.g. sale.csv. You must output to a CSV file with the name specified in <OUTPUT_CSV>. For example, the program could be run as:

```
SAMPLE RUN:
python HW1_streaming.py sale.csv output.csv
```

Evaluation: You can develop and test your code in a notebook using the sample file provided on NYU Classes. But **you must turn in a stand-alone script** that can be run through the command-line. We will run your code through a much larger data set. So please make sure that your code can handle the data in a streaming fashion.