MR JOIN HW, DUE: OCTOBER 10, 11:59PM EST

1. Instructions

1.1. **What to look at.** The only references you may use for this lab are:

- The mrjob api reference: `https://mrjob.readthedocs.io/en/latest/`
- Any material on Canvas

In particular, do not use stackoverflow or chegg or ChatGPT or MS Copilot for two reasons, they often give bad advice.

1.2. **How to collaborate.**

- Write down (in your code) the names of people you collaborate with.
- Cannot look at each other's code. You can collaborate by discussing ideas.

1.3. **Getting the code and submitting the code.** Use update-starter to get the code and submit to gradescope. You can submit multiple times based on autograder responses **Make sure to only commit human-generated files, and no other files into your repository.**

After running update-starter, you should see an **debuglab** directory for this assignment.

1.4. **The Assignment.**

Question 1. This question uses the Orders dataset in the HDFS **/datasets/orders** directory. For this question, write a mapreduce job to compute, for every combination of country and month, the total quantity of items bought in that country. For example, if `[Wakanda, 12] 99` appears in the output, it means that in December, all customers from Wakanda together bought 99 items. Make sure to fill in **q1/q1.py**, **q1/run**, **q1/testdata/orders.txt**, **q1/testdata/customers.txt**, **q1/testdata/items.txt**, and **q1/expectedoutput.txt**

Question 2. This question uses the Orders dataset in the HDFS **/datasets/orders** directory. For this question, write a mapreduce job to compute, for every item (StockCode), the number of **distinct** countries that bought this item. For example, if `IT9999 2` appears in the output, it means that the item with stock code IT9999 was only bought in 2 different countries. Make sure to fill in **q2/q2.py**, **q2/run**, **q2/testdata/orders.txt**, **q2/testdata/customers.txt**, **q2/testdata/items.txt**, and **q2/expectedoutput.txt**

Question 3. This question uses the Orders dataset in the HDFS **/datasets/orders** directory. Recall that each line of the orders chunk is a *part* of an order. All parts of the same order have the same order id (InvoiceNo). We are interested in how many orders have 1 part, how many orders have 2 parts, etc. So in the output of the mapreduce job, the key should be the number of parts and the value is the number of orders having that many parts. For example, if an output line is `4 56789`, it means there are 56789 orders that have 4 parts. Make sure to fill in **q3/q3.py**, **q3/run**, **q3/testdata/orders.txt**, **q3/testdata/customers.txt**, **q3/testdata/items.txt**, and **q3/expectedoutput.txt**