**Assignment 1, Big Data, Spring 2018, Due: February 5, 2019 by 11:59PM**

**Parts 1 and 2 are Individual Category C. Part 3 is Individual Category A.**

Part 1 and Part 2: Use DataA1.csv as the input. Data overview: The data is in a CSV file. The rows are Orders. The first column is Customer ID Number, the second column is Item ID Number, and the third column is Amount Spent on the Order. Note, that Customer ID Number and Item ID Number repeat (i.e., a customer will have multiple orders, and an item will be ordered more than once).

**Part 1:**

Use Python to create a MapReduce program to determine the Total Amount spent by Customer. Execute. Submit your code (.py file) and output (.txt file). Put your execution statement as the final line in your code (# commented out).

Hints:

Converting a string (stringtonumber) to a number (floating point): float(stringtonumber)

Adding up numbers coming into a reducer (value): sum(value)

**Part 2:**

Use Python to create a MapReduce program to sort the Total Amount spent by Customer (from the Customer who spent the least to the Customer who spent the most). Execute. Submit your code (.py file) and output (.txt file). Put your execution statement as the final line in your code (# commented out).

Hints:

You will probably need to chain two MapReduce jobs together.

Code Snippet: **'%04.02f'%float(order)**

Allows you to change a variable that is a float(order) to give it leading zeros. Where “order” was a float.

**Part 3:**

Find your own data set. Come up with a “business problem” to address using MapReduce, and complete it. Provide (~half page of text) an Executive Summary (that describes the data set, the business problem, what you did to solve it, and the overall answer); you may include a figure – but don’t exceed 1 page total. Include your code, output, and a link to the data set (if the data set is less than 2MB, then feel free to include it directly).

**Submission:**

Please put all of your files into a single folder, ZIP it, and then submit via Blackboard. **Save your ZIP file as LASTNAME-FIRSTNAME.ZIP.** (If you have issues uploading to Blackboard, then email it to me directly. Email **both** my Mason and GMAIL.)

**Emails:** Joe.Wilck@mason.wm.edu Joe.Wilck@gmail.com

**Grade:** Parts 1-2: 40% each. Part 3: 20%, but will be graded “tough” and level of difficulty, creativity, uniqueness, etc. will factor into the grade. Note, Assignment 1 will count as 20% of your overall course grade.

If you are looking for some places to find a data set, here is just a short list (in no particular order). Note that some of these are commercialized and some require more work/effort than others to extract data:

Note, some of these links are old (may be dead), but I think most are still working.

Stanford's Large Network Dataset Collection:

<http://snap.stanford.edu/data>

Yahoo Webscope Program:

<http://webscope.sandbox.yahoo.com/>

Text Retrieval Conference:

<http://trec.nist.gov/data.html>

US Government's Open Data:

<https://www.data.gov/>

Google’s Public Data Directory:

<https://www.google.com/publicdata/directory>

Amazon’s AWS:

<https://aws.amazon.com/public-datasets/>

Kaggle:

<https://www.kaggle.com/datasets>

Airbnb:

<http://insideairbnb.com/get-the-data.html>

MLB Pitch-by-Pitch:

<https://cloud.google.com/bigquery/public-data/baseball>

Peter Skomoroch (formerly LinkedIn) Data Links:

<https://del.icio.us/pskomoroch/dataset>

Wikipedia Database:

<https://en.wikipedia.org/wiki/Wikipedia:Database_download>

IMDb Database:

<http://www.imdb.com/interfaces>

Last.fm (music) Database:

<http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/index.html>

Quandl:

<https://www.quandl.com/>

Awesome Public Datasets:

<https://github.com/caesar0301/awesome-public-datasets>

Metro Boston:

<http://metroboston.datacommon.org/>

Census:

<http://www.census.gov/>

Dataverse:

<http://dataverse.org/>

Infochimps:

<http://www.infochimps.com/>

Linked Data:

<http://linkeddata.org/>

Reddit Open Data:

<https://www.reddit.com/r/opendata/>

Real Climate:

<http://www.realclimate.org/index.php/data-sources/>

Climate Station Files:

<http://www.metoffice.gov.uk/research/climate/climate-monitoring/land-and-atmosphere/surface-station-records>

CDC:

<https://www.cdc.gov/nchs/index.htm>

World Bank:

<http://datacatalog.worldbank.org/>

Maps (SDC – useful for creating graphics):

<http://www.d-maps.com/index.php?lang=en>

UK Office of National Statistics:

<https://www.ons.gov.uk/>

State Master:

<http://www.statemaster.com/index.php>