CPSC 4330 Big Data Analytics
Winter 2025

Homework 1

Due: 10:55 am, Friday, Jan. 24

In this homework, you will write your own MapReduce application to analyze customer reviews.

The file format is CSV, which uses comma to separate the fields in each row. Each line of the file represents one data record. We'll primarily be interested in the id of a product (the 1st column) and the rating of a product (the 3rd column). The sample file (Gift_Cards.csv) can be downloaded from Canvas.

The MapReduce application you should write will compute the total number of reviews and the average rate of each product. Note that each review record has a rating in it (i.e. the number of ratings = the number of reviews). The output should contain three columns (product_id, total number of reviews of this product, the average rate of this product).

You can test your code on the sample file (Gift_Cards.csv). If that looks good, you can upload your application to AWS, scale out worker nodes and run on a slightly larger data set (review_data.zip) which is available on Canvas. If you want to try on an even larger data set, there is also one (review_data_large.zip which has about 4.7G files) on Canvas.

Review exercise 4 on how to use AWS to run your mapreduce code. When your job is done, check the results to make sure everything looks good. Also, **remember to make sure that the cluster is terminated when you're done!**

Submission

Use Canvas to submit the following: (You can compress the files if that's convenient)

- All your source code files (.java)
  - Only include source files, not compiled files (i.e. no .class or .jar files)
- All your output files of running your code against the files in review_data.zip on AWS.