CPSC 4330 Big Data Analytics
Winter 2021

Homework 1

Due: 6:00 pm, Monday, Jan. 25th

In this homework, you will write your own MapReduce application to analyze Amazon customer reviews. Amazon has made a set of review data available publicly. Information about this data set can be found here: https://s3.amazonaws.com/amazon-reviews-pds/readme.html. Details about the structure of the data and different files in the data set can be found here: https://s3.amazonaws.com/amazon-reviews-pds/tsv/index.txt

We'll primarily be interested in the "star_rating" (the rating of a product) and "product_id" (the id of a product) columns of the data set. The file format is Tab Separated Values (TSV), which uses tab characters to separate the fields in each row. Each line of the file represents one data record, except that the first row contains headers (i.e. the names of the fields). The sample file (sample_us.tsv) is pre-loaded on VM, in /home/cloudera/training_materials/sampledata.

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 (sample_us.tsv). If that looks good, you can get one actual data set file to test on. Here is a file https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Electronics_v1_00.tsv.gz (about 667 MB). You can download it from the URL with the web browser, or you can use wget on the command line on VM to download it:

```
wget https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Electronics_v1_00.tsv.gz
```

This file has .gz on the end because it is a compressed file (containing a .tsv file). Hadoop automatically detects and handles gzip compressed files, so you can directly use it as input.

Now you can upload your application to AWS, scale out worker nodes and run on some larger data sets in https://s3.console.aws.amazon.com/s3/buckets/amazon-reviews-pds/tsv/ (after you log in to AWS, by typing this URL, you'll be able to see all the files in the tsv folder). To analyze the entire dataset in that folder, you need to start about 10 worker nodes (but it takes a really long time to start the cluster). For this homework, you can just analyze the reviews in these three files (in total 5.3G)

- amazon_reviews_us_Books_v1_00.tsv.gz (2.6G)
- amazon_reviews_us_Books_v1_01.tsv.gz (2.5G)
- amazon_reviews_us_Books_v1_02.tsv.gz (1.2G)

Review exercise 4 on how to use AWS to run your mapreduce code, but make the following changes:

- For the input argument of your job, you can point it to the S3 bucket where data files are already made publicly available:
  - s3://amazon-reviews-pds/tsv/amazon_reviews_us_Books_v1_0*.tsv.gz
  - Note the * wildcard means all .tsv.gz files that start with amazon_reviews_us_Books_v1_0. In the tsv folder, there are three files that match this pattern (i.e. 00, 01, 02).
- For the output argument, you'll still want to point it to your bucket, but you may want to name a different directory (so it does not collide with your previous directory in exercise 4).
- On the "Step 2: Hardware" page, you can increase the number of worker nodes (e.g. changing to 4 worker nodes should be sufficient for the workload)
- 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!**

<u>Submission</u>

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 on the three review files related to books
  - Be sure you submit the output of running on the three review files as input – don't submit the output from testing on a single data file by mistake.