Database Implementation System - Programming Assignment - Hadoop and MapReduce From Mrinal Monga

- 1. The tutorial provided in the Sakai resources was followed.
- 2. The MapReduce task that returns the average rating of movies in the Netflix dataset was implemented. Please see below for more details

Appropriate screenshots are given at the end of the document which make it easier to understand this report

Attachment contain the java programs and other files.

Submission related details -

- 1. Path to the s3 bucket with output s3://rutgers.monga.mrinal/
- For the Word Count example the output is at <u>s3://rutgers.monga.mrinal/</u>WordCountTestOutput and the cluster used was WordCountTest
- 3. For the Netflix Small dataset the output is at <u>s3://rutgers.monga.mrinal/Netflix_small_output</u> and the cluster used was NetflixS For this part the jar file used was movieratingaveragecalculator.jar which is located at <u>s3://rutgers.monga.mrinal/movieratingaveragecalculator.jar</u> The Java program which was used for this is the MovieRatingAverageCalculator.java program. The detailed description is given below.

MovieRatingAverageCalculator.java program description

- 1. **The map method** This method takes the second column of the input file as the id of the movie and the third column as the rating of the movie. Exception handling is done to take care of the situation when parsing a string from the tokenizer might not return a float value. The ratings are stored as float. This helps later during average calculation.
- 2. **The reduce method -** This method has two variables of float type sum and counter. sum stores the sum of all the ratings for a movie id. The counter stores the number of ratings. The average rating is arrived at by the formula sum/average for each movie id.
- 4. For the Netflix Large dataset the output is at
 - <u>s3://rutgers.monga.mrinal/Netflix_large_output</u> the cluster used was **NetflixL**. For this the same jar file was used as in for Netflix_small.
- 5. For sorting the movie average ratings the program used was SortingRating.java and the jar used was sortingrating.jar located at

s3://rutgers.monga.mrinal/sortingrating.jar

Output is at - <u>s3://rutgers.monga.mrinal/</u>SortedOutput the cluster used was **Netflix LARGE Sorting Output**

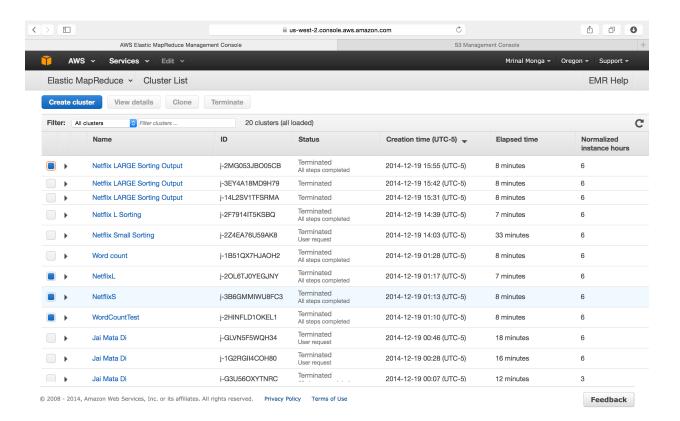
SortingRating.java

- 1. The main method first takes the input provided and in a temporary folder writes the movie ids and the average ratings for the movies. Then it takes the input from the temporary folder and produces the required sorted output.
- 2. The SortingRating.java has Map1, Reduce1, Map, Reduce, ReverseComparator

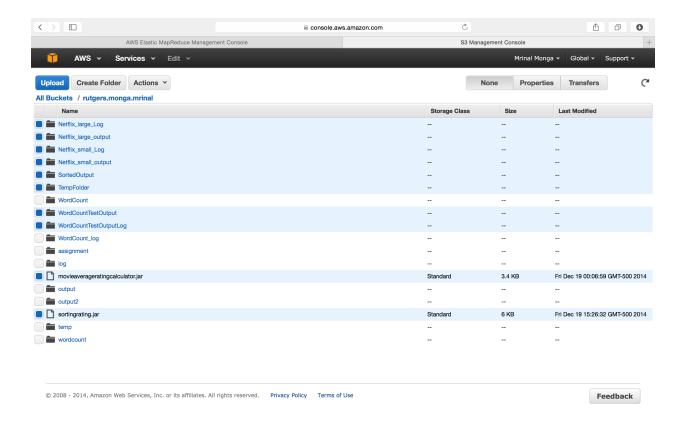
classes.

- 3. In the main method of the SortingRating.java the first job is of calculating the average rating. For this the Map1 and Reduce1 classes are used. The map and reduce methods of these classes do the same job as that of the map and reduce methods of the MovieAverageRatingCalculator.java
- 4. The second job performed in the main method of the SortingRating.java is sorting. For this the Reverse Comparator class is used. This causes the data to be arranged in descending order of average movie rating.

SCREENSHOT 1 - CLUSTERS



SCREENSHOT - BUCKET AND DIRECTORIES



PERMISSIONS

