Hadoop Assignment

Introduction

You need to write a Hadoop job that is able to aggregate and summarize movie ratings from a movie rating database.

There are two distinct phases of this project. Phase 1 will calculate averages ratings by movie. Phase 2 will take average movie ratings from Phase 1 to determine the distribution of ratings in the dataset.

Phase 1: Rating Aggregation

The goal of Phase 1 will be to aggregate movie reviews for each individual movie. For example, if a movie in the ratings database contains ten ratings, the result of this MapReduce iteration will be a single averaged rating for this movie.

This job will read from a file containing movie ratings in the following format:

User_id::Movie_id::rating

The **User_id** and **Movie_id** fields uniquely identify user and movies respectively. The rating field is a double value in the range [0.0 to 5.0] that represents a user's rating of the movie. Each entry is separated by a newline. An extraction from the file can be seen below.

User_3687::Movie_936::5.0
User_8107::Movie_211::0.5
User_6600::Movie_323::3.0
User_2997::Movie_367::0.5
User_3424::Movie_246::2.0

A file titled **Movie_Ratings.txt** in the above format will be used as input for Phase 1. The file will be split by lines and sent to the Map method, which will then parse the lines and emit <Movie_id, rating> pairs. The Reduce function will take these pairs and compute an average rating for each Movie_id.

Results from Phase 1 (i.e. <Movie_id, average-rating> pairs) should be stored in a file called average-ratings.txt. The file should contain one
<Movie_id, average-rating> pair per line.

Phase 2: Rating Summarization

The goal of Phase 2 is to take the average movie ratings that were output from Phase 1. A bin is just a range of values with a count that represents the total number of movies in the bin. A movie belongs to a bin if its average rating from Phase 1 falls into the bin's range. Because the bin ranges are mutually exclusive, each movie will belong to exactly one bin. There will be ten bins total, and there ranges will be as follows:

bin 1: [0.0, 1.0)

bin 2: [1.0, 2.0)

bin 3: [2.0, 3.0)

bin 4: [3.0, 4.0)

bin 5: [4.0, 5.0)

Phase 2 will input the txt file average-ratings.txt that was the result of the calculations from Phase 1. The Map <Movie_id, average-rating> method will parse these pairs and determine the appropriate bin ID based on the average movie rating. For each <Movie_id, average-rating> pair inthe input file, a <bin_id, 1> pair will be emitted. The Reduce method will take these <bin_id, 1> pairs and calculate a sum of pairs for each bin_id. The final result of Phase 2 will be a count for each bin representing the number of movies mapped to that bin. Write these results to a file called ratings-summary.txt