Benjamin Arancibia

December 8, 2014

IS607 Week 14 Assignment

Map/Reduce

*Situation*

I want to create a page rank by hour and then look at what the size of the top 5 page rank content was returned. Using the Wikipedia log dump from 2007 until 2014 there are hourly statistics that are in the following format.

en Main\_Page 242332 4737756101

This format shows that the English language main page of Wikipedia was requested 242332 times (not unique visits) and 4737756101 is the size of the content returned.

*Map and Reduce*

In this step after downloading the data and unzipping it we will upload the data to MongoDB. If possible and there is an API available we use that API to gather the data, but for this situation we will be reading in the data. The mapper will read in a Wikipedia logs and the mapper will process the logs one at a time. For each record the mapper extracts the language, page, times visited and amount content returned.

The reducer will iterate over all the data and count the number of page views per hour and the associated amount of data per page. The reducer will then sort the data by the top 5 page rank (number of visits to a Wikipedia page). Based on this sort, the reducer will write the results of the top 5 page rank and the content returned.

*Important Aspects of Map/Reduce*

There are some important aspects of Map/Reduce:

* Map tasks should be stateless. Map tasks should calculate the same results strictly based on the input data, regardless of the order of the input data.
* Reduce tasks should be stateless but may maintain state while iterating through inputs with the same key.
* Tasks cannot communicate with one another. Each task should be able to commute its result based strictly on the input data.