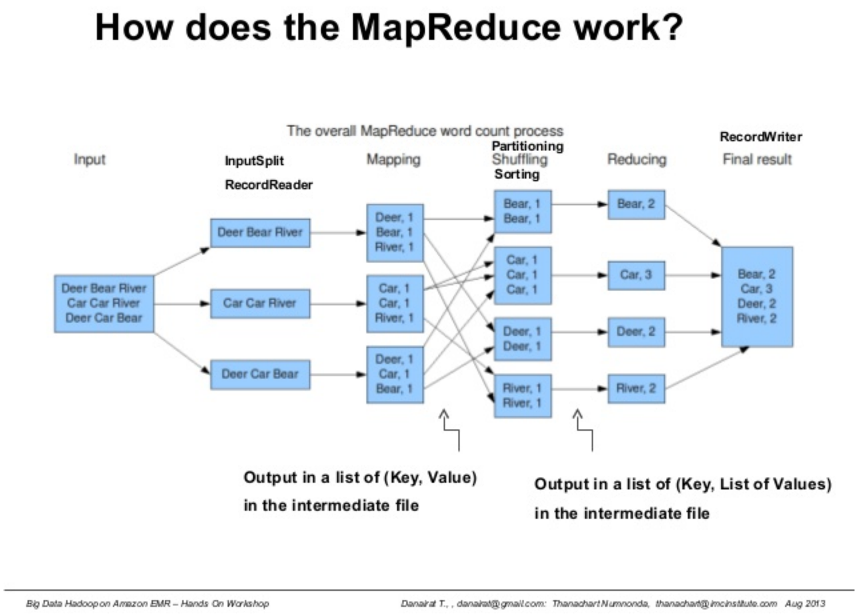
Cmpe 300 Fall 2017
Application Project

November 13, 2017

In this project, you are supposed to implement the MapReduce algorithm on MPI to count the word occurrences in a file. MapReduce is basically explained in the figure below [1].



How does the MapReduce work?

Big Data Hadoop on Amazon EMR – Hands On Workshop

Danairat T., , danairat@gmail.com: Thanachart Numnonda, thanachart@imcinstitute.com Aug 2013

First, input is partitioned into segments. Then each segment is sent to a distinct process and word occurrences are counted which corresponds to the mapping stage. Notice that in this step, there is no cumulative count, you just mark the words that occur. Later, the list is put together and sorted. Last step is reducing, which outputs the total count of each word.

Your program flow should be as follows:

1. Split the input and send them to slaves
2. Slaves map the words and send it back to master
3. Split the input again and send them to slaves to be sorted (Mergesort seems the most convenient to me although you can choose any sorting algorithm you like. Note that you will have to do the last merge in the master process, this is acceptable)
4. The master process reduces the list

Since you will use MPI with C for this project, defining your struct can be a good idea. A struct with a char array (of enough size for the longest words) to store the word and an integer to store the count value will suffice. This is just a tip, you are not limited on your implementation ideas.

Input and Output

You will be given an input file which contains a speech by a historical person. We will provide the tokenized version of this input file for your convenience. You can directly use this tokenized input. Your program should output the word counts on the terminal window or in a file.

As a side quest, you can try to guess the author of the speech.

Submission

You are supposed to submit a zip file that contains your source code and a report written in this format. There will be a demo session after projects are submitted. Demo dates and hours will be announced in the meantime.

References

[1] *https://www.slideshare.net/imcinstitute/big-data-hadoop-using-amazon-elastic-mapreduce-handson-labs*. Accessed at Nov 13, 2017.