Data Mining

Homework 2

Taking a census of the United States.

With nearly 330 million people living in the United States, taking a census every ten years can be a challenging and time-consuming task. MapReduce is a great solution offering many advantages. There are three steps in MapReduce. Map streamlines the data into a HashMap containing zip codes as keys and members in one individual household as values. Group groups all the values belonging to the same key. Reduce aggregates the total of the value belonging to each individual key. Using MapReduce, the data can be parallelized which saves memory and time, making it more efficient. Rather than calculating how many people live in the United States at once, MapReduce divides the task by area, so multiple algorithms can run at the same time, and then those areas can be grouped together to calculate the total.

Pseudo-code:

```
Read in CSV as data frame
Initialize HashMap
For row in data:
    Append <zip code, Members in Household> to HashMap
Group HashMap by zip code
Reduce HashMap to sum of household members for each zip code
Merge HashMap with a table of zip codes and states on zip codes
Group by state
Reduce HashMap to sum of total household members for each state
Reduce total sum
```