Homework 3: Writing Component

UNITED STATES CENSUS DATA ANALYSIS USING MAPREDUCE
VERSION 1.0

DUE DATE: Friday, April 14th, 2017 @ 5:00 pm [US Census Data Analysis]

Q1. How would you implement the complete set of tasks in HW3-PC using a single MapReduce job? Note that your program should emit outputs for each task into a separate file.

[300-400 words]

Q2. Suppose that the U.S. Census Bureau has introduced a new scheme to disseminate demographic data. In addition to releasing the entire dataset as a batch at the end of a census cycle, the Census Bureau is now providing changes to the population (number of new births or deaths for each census block for a particular time period) as a continuous stream of updates. How would you extend your solution to support this new scheme in order to provide a more up-to-date view of the U.S. population?

[400-500 words]

Q3. Instead of supporting a set of fixed queries as you have done in HW3-PC, you are being asked to implement a solution that can run any arbitrary query on the census dataset using the Hadoop infrastructure. How would you design your solution to cope with this requirement?

[400-500 words]

Q4. How would you extend your current solution to identify locations that are most similar to each other and also those that are most dissimilar to each other? Your similarity measures may include distributions relating to: age, gender, etc.

[400-500 words]

Q5. Consider the case where two additional datasets has been made available to you. The first dataset includes information about *migration patterns* (alongside demographics such as age, gender, and educational levels) into and out of a state. The second dataset includes *economic data* such as number of new jobs that were created, the sectors that these jobs were created in, the average pay, educational levels requirements for jobs, etc.

Describe how you will design a framework that allows you to make reasonably accurate projections about the expected population levels in a particular state. Assume that you have census, migration patterns, and economic data for 1990, 2000, and 2010. In the case of migration patterns and economic data assume that you also have yearly data for the past 10 years.

[~500 words]

1 Grading

Homework 3 accounts for 20 points towards your final course grade. This written component accounts for 20% of the points set aside for HW3 i.e. this assignment accounts for 4% of your cumulative course grade. This assignment is graded on a 20-point scale with each question accounting for 4points.

2 What to Submit

You should submit a PDF document. Please use the following naming convention: HW3-WC-Firstname-Lastname.pdf.

The folder set aside for this assignment's submission using checkin is **HW3-WC**