Homework 3: Programming Component

UNITED STATES CENSUS DATA ANALYSIS USING MAPREDUCE
VERSION 1.1

DUE DATE: Wednesday, April 12th, 2017 @ 5:00 pm

OBJECTIVE

As part of this assignment you will be working with datasets released by the United States Census Bureau. You will be developing MapReduce programs that parse and process the 1990 US Census dataset to support knowledge extraction over demographic data from all fifty states.

You will be using Apache Hadoop (version 2.7.3) to implement this assignment. Instructions for accessing datasets and setting up Hadoop clusters will be available on the course website.

This assignment may be modified to clarify any questions (and the version number incremented), but the crux of the assignment and the distribution of points will not change.

1 Cluster setup

As part of this assignment you are responsible for setting up your own Hadoop cluster with HDFS running on every node. We will be staging the 1990 US Census dataset on a *read-only* cluster. You should use your ***own*** cluster to write outputs produced by your MapReduce programs. MapReduce clients will be able to access namespaces of both clusters through Hadoop ViewFS federation. Your programs will process the staged datasets; data locality will be preserved by the MapReduce runtime.

2. Analysis of US Census Data

You will be working with a subset of the 1990 US census dataset that has been staged for you in the shared HFDS cluster. This dataset captures various population and housing information across all states in 1990. We are using the summary tape file 1(b) data from the entire Qdataset.

You need to process this dataset using MapReduce to answer the following questions.

Q1 On a per-state basis provide a breakdown of the percentage of residences that were rented vs. owned.

Q2. On a per-state basis what percentage of the population never married? Report this for both males and females. Note: The US Census data tracks this information for persons with ages 15 years and over.

Q3. On a per-state basis, analyze the age distribution (of the population that identifies themselves as Hispanic) based on gender.
(a). Percentage of people below 18 years (inclusive) old.
(b). Percentage of people between 19 (inclusive) and 29 (inclusive) years old.
(c). Percentage of people between 30 (inclusive) and 39 (inclusive) years old.

Q4. On a per-state basis, analyze the distribution of rural households vs. urban households.

Q5. On a per-state basis, what is the median value of the house that occupied by owners?

Q6. On a per-state basis, what is median rent paid by households?

Q7. What is the 95^{th} percentile of the average number of rooms per house across all states?

Q8. Which state has the highest percentage of elderly people (age > 85) in their population?

Q9. Come up with an innovative analysis program for this dataset. For this component, think of yourself as the lead data scientist at a start-up firm. What would do with this dataset that is cool?

You are allowed to: (1) combine your analysis with other datasets, (2) use other frameworks such as Mahout for performing your analyses, and/or (3) perform visual analytics.

Restrictions: Note that there should be NO DISCUSSIONS about Q9 on Piazza. Your analysis must be something that you have come up with on your own.

Q9 is quite open-ended and you have a lot of freedom. That freedom comes with the responsibility that you manage your own problems and don't except someone else (be it the Professor, GTA, or your peers) to solve your problems for you. You have to iron out all problems that you are facing on your own.

1.1.1 Dataset

The dataset comprises a collection of files. Each file has a set of records representing a series of data items that capture different demographic information. Each record is a continuous series of 9610 characters that you need to split at appropriate boundaries to extract relevant fields. A particular record is divided into 2 record segments of 4805 characters with each segment having 300 characters of identification information followed by tables. Each segment has a set of geographic information, that is encoded in the first 300 characters. The layout of these first 300 characters for each segment is identical. Each segment appears as a new line in an input file. The logical record has a record sequence number, which is repeated in each segment. The logical record number appears in position 19 (the starting index for a record is 1, not 0) of each segment. Following this, beginning in positions 25 and 29, are the logical record part number and the total number of parts in the record. By viewing these two fields together, the sequence of the segment and the total number of segments can be quickly determined. For example, 1 in the logical record part number and 2 in the total number of parts in record field indicates that this is segment 1 of the 2 segments that comprise the logical record. You should pay attention to the record part number field when trying to extract a particular field using the boundaries provided below, because a particular field appears in either segment 1 or 2. Boundaries in the following table are defined with respect to the starting position of the segment.

Each record is associated with a summary level, which is hierarchical. For instance, in summary level 140, the hierarchy listed is State—County—census tract/block numbering area. This record contains data for a census tract/block numbering area within a particular county of a state. The dataset contains records at different summary levels. Some of the records are in summary levels higher up in the hierarchy, which summarizes the information captured at lower summary levels. So you should be careful to process only the records at the appropriate summary levels to avoid duplicate processing. In this assignment, you will be working with the lowest level of summary data to answer the questions above. **You will ONLY be processing records at summary level of 100.**

The table below summarizes the all the fields necessary to implement your MapReduce program. It lists all the fields and their corresponding boundaries. Note that the boundary indices starting at 1.The complete documentation including the data dictionary for the dataset is available at http://www2.census.gov/census_1990/STF1B_ASCII/TechDoc/D1-D90-S100-14-TECH.pdf.

The dataset is available under directory `/data/census` in the shared HDFS.

Fields	Starting index	Fields Size	Segment	Data Type
State/US Abbreviation	9	2	1, 2	Alphanumeric
Summary Level	11	3	1, 2	Numeric
Logical Record Number	19	6	1,2	N
Logical Record Part Number	25	4	1,2	N
Total Number of Parts in Record	29	4	1,2	N
Population				
Persons	301	9	1	Numeric
Urban and Rural				
Inside urbanized area	328	9	1	Numeric
Outside urbanized area	337	9	1	Numeric
Sex				
Male	364	9	1	Numeric

Female	373	9	1	Numeric
Age				
Under 1 year	796	9	1	Numeric
1 and 2 years	805	9	1	Numeric
3 and 4 years	814	9	1	Numeric
5 years	823	9	1	Numeric
6 years	832	9	1	Numeric
7 to 9 years	841	9	1	Numeric
10 and 11 years	850	9	1	Numeric
12 and 13 years	859	9	1	Numeric
14 years	868	9	1	Numeric
15 years	877	9	1	Numeric
16 years	886	9	1	Numeric
17 years	895	9	1	Numeric
18 years	904	9	1	Numeric
19 years	913	9	1	Numeric
20 years	922	9	1	Numeric
21 years	931	9	1	Numeric
22 to 24 years	940	9	1	Numeric
25 to 29 years	949	9	1	Numeric
30 to 34 years	958	9	1	Numeric
35 to 39 years	967	9	1	Numeric
40 to 44 years	976	9	1	Numeric
45 to 49 years	985	9	1	Numeric
50 to 54 years	994	9	1	Numeric
55 to 59 years	1003	9	1	Numeric
60 and 61 years	1012	9	1	Numeric
62 to 64 years	1021	9	1	Numeric
65 to 69 years	1030	9	1	Numeric
70 to 74 years	1039	9	1	Numeric
75 to 79 years	1048	9	1	Numeric
80 to 84 years	1057	9	1	Numeric
85 years and over	1066	9	1	Numeric
Age by Gender [Hispanic Origin]				
Male: Under 1 year	3865	9	1	Numeric
Male: 1 and 2 years	3874	9	1	Numeric
Male: 3 and 4 years	3883	9	1	Numeric
Male: 5 years	3892	9	1	Numeric
Male: 6 years	3901	9	1	Numeric
Male: 7 to 9 years	3910	9	1	Numeric
Male: 10 and 11 years	3919	9	1	Numeric
Male: 12 and 13 years	3928	9	1	Numeric
Male: 14 years	3937	9	1	Numeric
Male: 15 years	3946	9	1	Numeric
Male: 16 years	3955	9	1	Numeric
Male: 17 years	3964	9	1	Numeric
Male: 18 years	3973	9	1	Numeric
Male: 19 years	3982	9	1	Numeric
Male: 20 years	3991	9	1	Numeric
Male: 21 years	4000	9	1	Numeric
Male: 22 to 24 years	4009	9	1	Numeric
Male: 25 to 29 years	4018	9	1	Numeric
Male: 30 to 34 years	4027	9	1	Numeric
Male: 35 to 39 years	4036	9	1	Numeric

Male: 40 to 44 years	4045	9	1	Numeric
Male: 45 to 49 years	4054	9	1	Numeric
Male: 50 to 54 years	4063	9	1	Numeric
Male: 55 to 59 years	4072	9	1	Numeric
Male: 60 and 61 years	4081	9	1	Numeric
Male: 62 to 64 years	4090	9	1	Numeric
Male: 65 to 69 years	4099	9	1	Numeric
Male: 70 to 74 years	4108	9	1	Numeric
Male: 75 to 79 years	4117	9	1	Numeric
Male: 80 to 84 years	4126	9	1	Numeric
Male: 85 years and over	4135	9	1	Numeric
Female: repeat age range	Starting from 4144	9 characters per each age range	1	Numeric
Gender by Marital Status (15 years and over)				
Male: Never Married	4423	9	1	Numeric
Male: Now Married, except separated	4432	9	1	Numeric
Male: Separated	4441	9	1	Numeric
Male: Widowed	4450	9	1	Numeric
Female: Repeat marital status	Starting from 4468	9 characters per each status	1	Numeric
Tenure				
Owner Occupied	1804	9	2	Numeric
Renter Occupied	1813	9	2	Numeric
Houses: Urban vs. Rural				
Urban: Inside urbanized area	1822	9	2	Numeric
Urban: Outside urbanized area	1831	9	2	Numeric
Rural	1840	9	2	Numeric
Not defined for this file	1849	9	2	Numeric
Houses: Rooms				
1 room	2389	9	2	Numeric
2 rooms	2398	9	2	Numeric
3 rooms	2407	9	2	Numeric
4 rooms	2416	9	2	Numeric
5 rooms	2425	9	2	Numeric
6 rooms	2434	9	2	Numeric
7 rooms	2443	9	2	Numeric
8 rooms	2452	9	2	Numeric
9 rooms	2461	9	2	Numeric
Value: Specified owner-occupied				
Less than $15,000	2929	9	2	Numeric
$15,000 - $19,999	2938	9	2	Numeric
$20,000 - $24,999	2947	9	2	Numeric
$25,000 - $29,999	2956	9	2	Numeric
$30,000 - $34,999	2965	9	2	Numeric
$35,000 - $39,999	2974	9	2	Numeric
$40,000 - $44,999	2983	9	2	Numeric
$45,000 - $49,999	2992	9	2	Numeric
$50,000 - $59,999	3001	9	2	Numeric
$60,000 - $74,999	3010	9	2	Numeric
$75,000 - $99,999	3019	9	2	Numeric

$100,000 - $124,999	3028	9	2	Numeric
$125,000 - $149,999	3037	9	2	Numeric
$150,000 - $174,999	3046	9	2	Numeric
$175,000 - $199,999	3055	9	2	Numeric
$200,000 - $249,999	3064	9	2	Numeric
$250,000 - $299,999	3073	9	2	Numeric
$300,000 - $399,999	3082	9	2	Numeric
$400,000 - $499,999	3091	9	2	Numeric
$500,000 or more	3100	9	2	Numeric
Contract Rent				
Less than $100	3451	9	2	Numeric
$100 to $149	3460	9	2	Numeric
$150 to $199	3469	9	2	Numeric
$200 to $249	3478	9	2	Numeric
$250 to $299	3487	9	2	Numeric
$300 to $349	3496	9	2	Numeric
$350 to $399	3505	9	2	Numeric
$400 to $449	3514	9	2	Numeric
$450 to $499	3523	9	2	Numeric
$500 to $549	3532	9	2	Numeric
$550 to $ 599	3541	9	2	Numeric
$600 to $649	3550	9	2	Numeric
$650 to $699	3559	9	2	Numeric
$700 to $749	3568	9	2	Numeric
$750 to $999	3577	9	2	Numeric
$1000 or more	3586	9	2	Numeric
No cash rent	3595	9	2	Numeric

2 Provided Resources

Datasets required for both components are shared through a viewfs based federated HDFS setup running on CS department machines. A complete guide on setting up your own Hadoop cluster and connecting to the shared HDFS has been provided on the course website.

3 Grading

Homework 3 accounts for 20 points towards your final course grade. The programming component accounts for 80% of these points with the written element (to be posted later) accounting for the remaining 20%. This programming assignment will be graded for 16 points. The point distribution for this assignment is listed below.

2 points	For setting up the Hadoop cluster
11 points	Knowledge extraction and developing programs to answer questions Q1 through Q8. You will also be judged on the elegance of your MapReduce programs. While getting the answers is important, your design matters as well.
3 points	Your solution to Q9

The grading for this assignment will be done based on a one-on-one interview and will include a code review.

4 Milestones:

You have 5 weeks to complete this assignment. The weekly milestones below correspond to what you should be able to complete at the end of every week.

Milestone 1: You should be able to set up a Hadoop cluster and get started with basic processing for the US Census data analysis.

Milestone 2: Programs to answer Q1, Q2, and Q4 are completed. Come up with the core idea for Q9.

Milestone 3: Programs to answer Q3, Q5, and Q6 are complete. Work on Q9 underway with significant progress.

Milestone 4: Programs to answer Q7, Q8, and Q9 are complete. Iron out bugs in any of the other components.

5 What to Submit

Use the CS455 checkin program to submit a single .tar file that contains:

- All the Java files related to the assignment (please document your code)
- You should use **`ant`** to compile your codebase and provide the corresponding `build.xml` file that is used for compiling the codebase. Please make sure that your `build.xml` works! You may modify the sample build.xml file that we have provided to do this.
- A `README.txt` file containing a description of each file and any information you feel the GTA needs to grade your program.

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

6 Change History

This section will reflect any changes that were made to a particular version of the assignment. Generally, these changes are made to better clarify the spirit of the assignment.

Version	Date	Change
1.1	3/9/2017	First public release of the assignment