MAPREDUCE IMPLEMENTATION

BY

AAROHI PATEL (1001452956)

VINAYAK TARE (1001453869)

SUBJECT
CSE 5331: DBMS MODELS AND IMPLEMENTATION
PROF. SHARMA CHAKRAVARTHY

STATUS

We have successfully developed the solutions for the two given mapreduce tasks. Also, for the second problem, we have built few histograms in excel for plotting age-wise frequency for each year.

IMPLEMENTATION

1. **Computing the average salary of males and females for each state for the 5 years. Compare them.**

Mapper:

```
Input key=Line number

Input value=csv content

Output key=State,sex

Output value=salary
```

In the mapper, for each line of csv file, we parsed the concatenation of state code along with the sex code as key and the salary value as the value.

Reducer:

```
Input key=State,sex

Input value=salary

Output key=State,sex

Output value=averageSalary
```

In the reducer, we iterated through the values for each key and calculated the sum of salaries for each gender for each state. Also for each key, we maintained a count to retrieve the number of individuals in each class. In the end we parsed the same key obtained from the mapper as the key for the reducer and totalSalary/Count i.e the average salary for each group as the value for the reducer.

2. **Equi Width Histogram on Age: use the buckets 0 to 9, 10 to 19, …., 90 to 99 for each year of data given and compare them.**

Mapper:

```
Input key=Line number

Input value=csv content

Output key=Year,AgeRange

Output value=”1”
```

In the mapper, for each line of csv file, we parsed the concatenation of year code along with the age range code as key and “1” as value. The age ranges are obtained by running and if else condition on the age values.

Reducer:

```
Input key=Year,AgeRange

Input value=”1”

Output key=Year,AgeRange

Output value=totalCount
```

In the reducer, we iterated through the values for each key and calculated the total individual in each age range for each year. In the end we parsed the same key obtained from the mapper as the key for the reducer and Count i.e the total number of people in each group as the value for the reducer.

Then, for each year we built a histogram in excel, indicating the age range bins as X-axis and frequency as Y axis.

RESULT

1. **Computing the average salary of males and females for each state for the 5 years. Compare them.**

01.1	45283.22748154072
01.2	27956.066570020295
02.1	45947.18175205238
02.2	31859.552006133403
04.1	47212.14492362134
04.2	32541.07576587648
05.1	40148.42720085003
05.2	26571.111301969016
06.1	55388.81676167811
06.2	38830.0178634363
08.1	53338.188661016204
08.2	34494.89728688393
09.1	70674.96842105263
09.2	41869.792647090835
10.1	49712.37757078299
10.2	35954.86055776892
11.1	73495.70814923907
11.2	56252.644988266846
12.1	48010.545963943856
12.2	32154.925151591415
13.1	49583.09007056632
13.2	32357.720717630153
15.1	46799.57620030112
15.2	33749.21818293658
42.1	47778.87384554189
42.2	30916.955295400978
44.1	50999.46260997067
44.2	34961.66017953322
45.1	43106.65245382974
45.2	28205.410545576757

```
46.1    37247.91254968768
46.2    24523.712047387027
47.1    44912.14335210047
47.2    29155.16463694693
48.1    51344.81589590177
48.2    32216.51473455641
49.1    47908.24604904632
49.2    25483.961001125605
50.1    41855.67270145545
50.2    29589.566225933657
51.1    59243.47240829941
51.2    38659.64213187942
53.1    54100.407217846274
53.2    34597.9453867285
54.1    42179.73044571208
54.2    26413.021688006273
55.1    43620.05779334501
55.2    29224.163826189644
56.1    46829.66352123168
56.2    27490.461082910322
```

Here, 0-56 represents state code and 1 and 2 are sex codes

1: male

2: female

And the 3rd value is average salary.

2. **Equi Width Histogram on Age: use the buckets 0 to 9, 10 to 19, …., 90 to 99 for each year of data given and compare them.**

Year AgeRange   ,Count

```
2009 0-9        ,203122
2009 10-19      ,221903
2009 20-29      ,192475
2009 30-39      ,200214
2009 40-49      ,233589
```

2009 50-59	,237565
2009 60-69	,178242
2009 70-79	,108396
2009 80-89	,58087
2009 90-99	,11467

2010 0-9	,203621
2010 10-19	,221475
2010 20-29	,197869
2010 30-39	,202841
2010 40-49	,232192
2010 50-59	,241236
2010 60-69	,185657
2010 70-79	,109956
2010 80-89	,60141
2010 90-99	,12139

2011 0-9	,196596
2011 10-19	,225720
2011 20-29	,205383
2011 30-39	,196234
2011 40-49	,228052
2011 50-59	,250141
2011 60-69	,197352
2011 70-79	,116558
2011 80-89	,64577
2011 90-99	,14982

2012 0-9	,196697
2012 10-19	,223104
2012 20-29	,203714
2012 30-39	,197118
2012 40-49	,226024
2012 50-59	,250805
2012 60-69	,204112

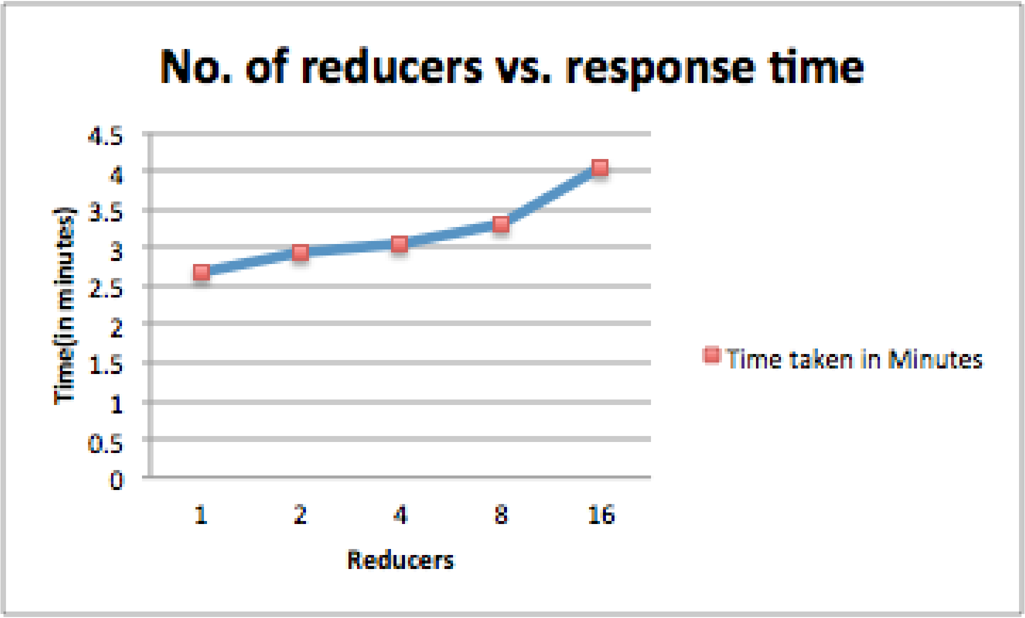
2012 70-79	,119263
2012 80-89	,63597
2012 90-99	,14790

2013 0-9	,196561
2013 10-19	,224066
2013 20-29	,206101
2013 30-39	,202643
2013 40-49	,222729
2013 50-59	,252746
2013 60-69	,208560
2013 70-79	,121886
2013 80-89	,62040
2013 90-99	,14830

PERFORMANCE ANALYSIS

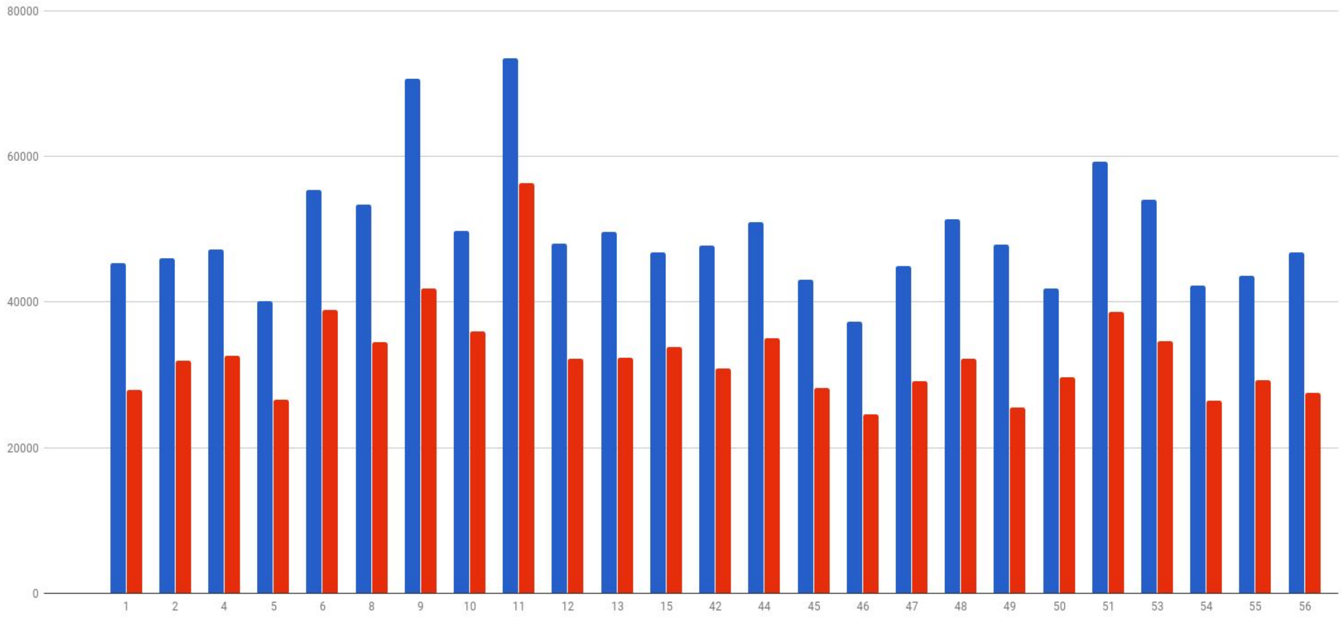
Number of Reducers	Time taken in Minutes
1	2.672123
2	2.925173
4	3.03749
8	3.322541
16	4.0299

Increase in the number of reducers increases the processing time. This might be because the time to start all reducers would be more, because of increase in number of reducers. Also the splitting of the data into more number of splits would increase the splitting time.



RESULT ANALYSIS

1. Computing the average salary of males and females for each state for the 5 years. Compare them.



This figure shows average salary of males and females of different states. Here 11 represents District of Columbia/DC has the highest salary whereas 46 which represents South Dakota has lowest average salary.

For all the states average salary of males has remained higher than the females in all these years.

2. Equi Width Histogram on Age: use the buckets 0 to 9, 10 to 19, …., 90 to 99 for each year of data given and compare them.

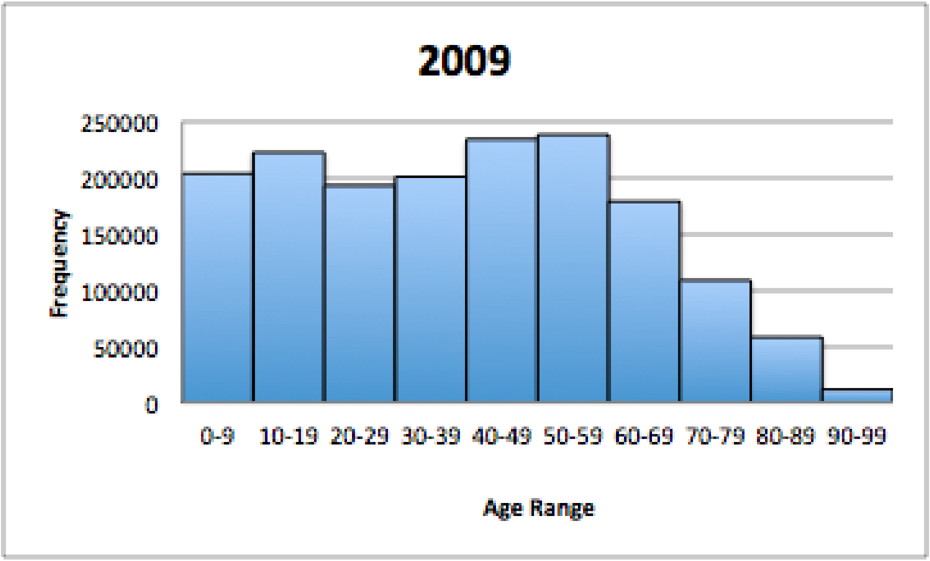


Fig. Histogram for Frequency for different age ranges for the year 2009.

We can see that in 2009, the highest number of people are in the range 40-60.

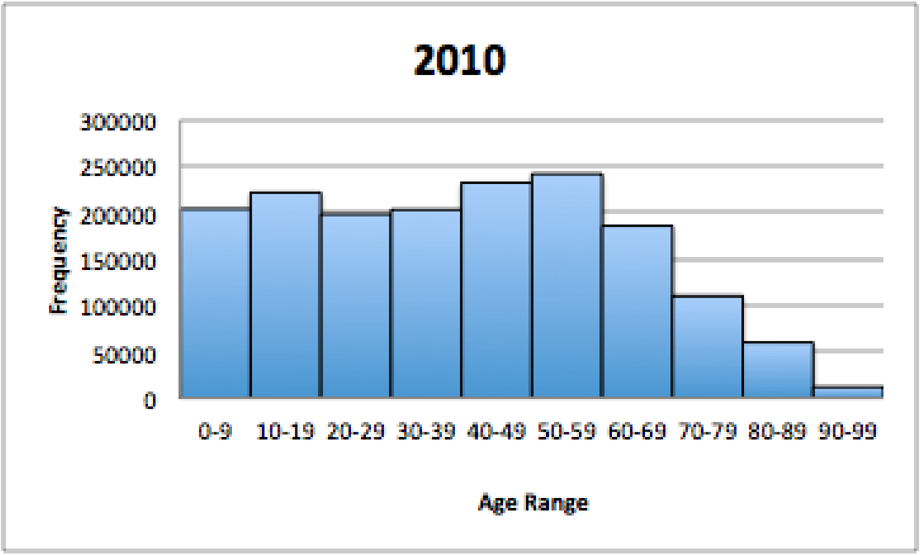


Fig. Histogram for Frequency for different age ranges for the year 2010

This shows that there is no major change in the frequency from 2009 to 2010.

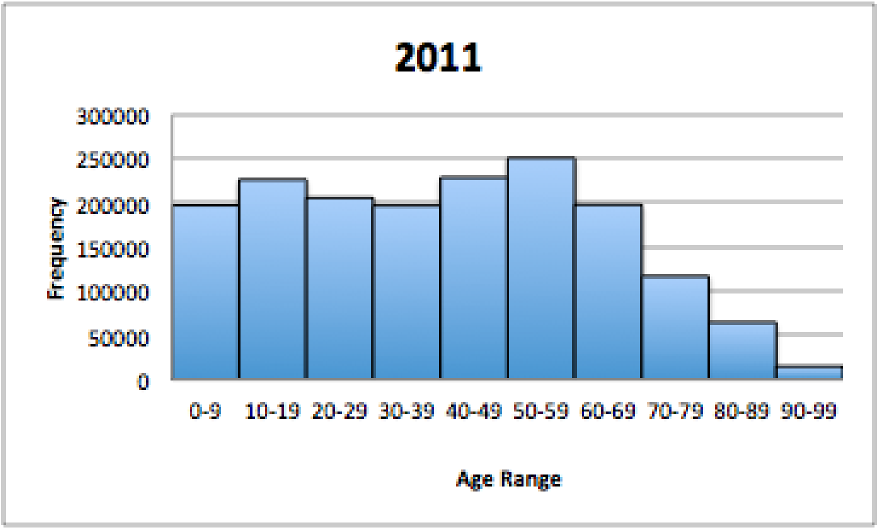


Fig. Histogram for Frequency for different age ranges for the year 2011

Compared to the previous year, the population in the age range 40-49 reduced and 60-69 increased drastically. Rest were almost similar.

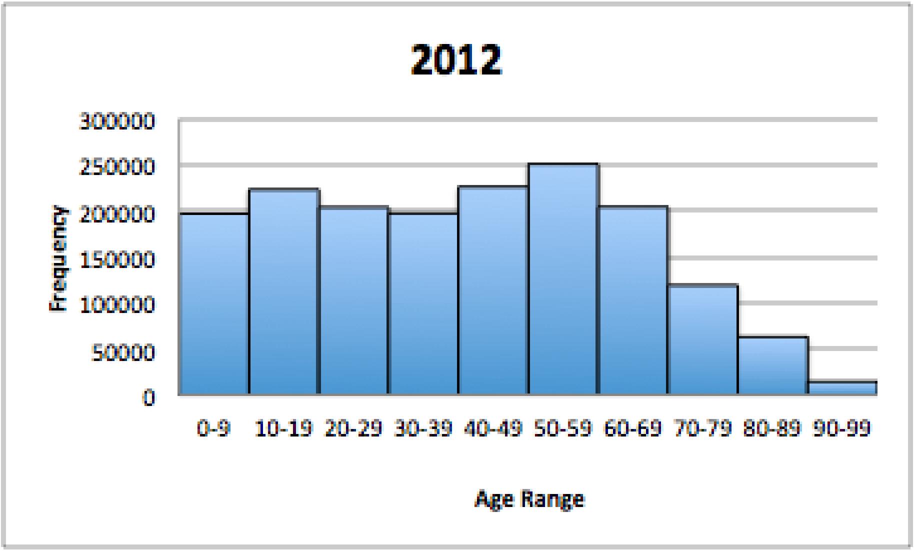


Fig. Histogram for Frequency for different age ranges for the year 2012

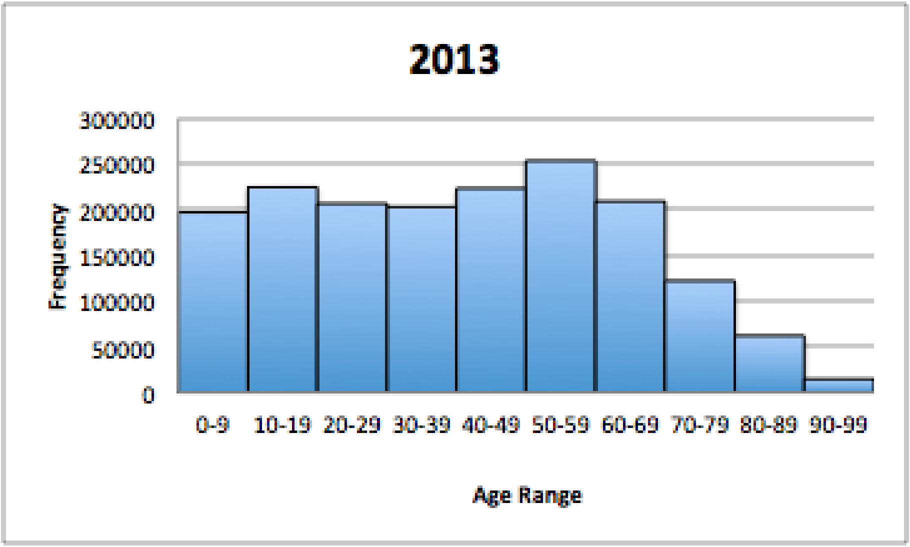


Fig. Histogram for Frequency for different age ranges for the year 2013

Number of people in age range 40-49 have decreased from the year 2009-2013. Overall, the population in the age range 40-49 has decreased and 60-69 has increased.

Apart from these deviation the percentage of people under each age group has remained almost same through all these years.

FILE DESCRIPTION

The Projects consist of 2 jar files with the name AvgSalary.jar and Histogram.jar.

RESPONSIBILITIES

TASK	PERFORMED BY	TIME
Installation	Vinayak Tare Aarohi Patel	5
MapReduce 1	Aarohi Patel Vinayak Tare	2
Multiple MapReduce	Vinayak Tare	3
MapReduce 2	Aarohi Patel	1
Histogram Plot	Vinayak Tare	1
Analysis	Aarohi Patel Vinayak Tare	3
Project Report	Vinayak Tare Aarohi Patel	3
Total Hours		18

LOGICAL ERRORS AND SOLUTIONS

1. While considering the average salary in mapreduce, we were taking int data type for summation of all salaries. However, results were incorrect because the total salary was very large as compared to the size of the int. We realized this problem and converted it to Long to get the correct results.
2. Was unable to add 2 keys separated by a comma. So used concat to merge the keys in a string.
3. .setNumMapTask() was unavailable in our version.