BIG DATA –Spring 2019

ASSIGNMENT 2

Due Date: 24th Feb 2019 10:00 PM on google classroom. Upload the Source code and the output file on Google Classroom.

Assignment Data Set: Consider the given bank dataset for both assignment questions. The dataset consists of 17 attributes. The attributes are separated by semicolon and are in following order

```
age;"job";"marital";"education";"default";"balance";"housing";"loan";"contact";"day";"month";"durat
ion";"campaign";"pdays";"previous";"poutcome";"y"
```

Further detail of the attributes is given in the dataReadMe file.

Question: ASSOCIATIVE MEMORY AND COMBINER

Two of the attributes in the given dataset are: age and job. For each **age**, find the maximum, minimum and average account balance of the clients.

You must provide the code for Mapper, Reducer and Combiner. You have to use **associative memory (array) in Mapper** to make your program efficient.

A portion of Input File

```
30;"unemployed";"married";"primary";"no";1787;"no";"no";"cellular";19;"oct";79;1;-1;0;"unknown";"no"
33;"services";"married";"secondary";"no";4789;"yes";"yes";"cellular";11;"may";220;1;339;4;"failure";"no"
35;"management";"single";"tertiary";"no";1350;"yes";"no";"cellular";16;"apr";185;1;330;1;"failure";"no"
30;"management";"married";"tertiary";"no";1476;"yes";"yes";"unknown";3;"jun";199;4;-1;0;"unknown";"no"
35;"management";"single";"tertiary";"no";747;"no";"no";"cellular";23;"feb";141;2;176;3;"failure";"no"
36;"unemployed";"married";"tertiary";"no";307;"yes";"no";"cellular";14;"may";341;1;330;2;"other";"no"
```

You Output for above portion of input file will be

Age = 30, Max Balance = 1787, Min Balance = 1476, Avg Balance = 1631
Age = 33, Max Balance = 4789, Min Balance = 4789, Avg Balance = 4789
Age = 35, Max Balance = 1350, Min Balance = 747, Avg Balance = 1048
Age = 36, Max Balance = 307, Min Balance = 307, Avg Balance = 307

Hint: In a Mapper Class, declare an array, override setup function for initializing array and override cleanup function to emit values.

Read Book Hadoop definitive guide for more details on how to use setup and cleanup methods of mapper class and reducer class (Pg 225, 243, 271)

Also have a look at map reduce documentation
https://hadoop.apache.org/docs/r1.2.1/api/org/apache/hadoop/mapred/Mapper.html
https://hadoop.apache.org/docs/current/api/org/apache/hadoop/mapreduce/Reducer.html

```
@Override
protected void setup(Context context) throws IOException, InterruptedException {
  //To initialize your associative array
}
```

```
@
Override
protected void cleanup(Context context)throws IOException, InterruptedException {
  ///do your work here
}
```

Question 2: PARTITIONER

We want to sort and partitioned the given dataset into different files based on the client's job. We wish to analyze clients doing different jobs separately.

You program will take the dataset and create a separate file for each job type (containing all the employees of that job type).

Hints: (consult book Hadoop the definitive guide esp. Pg220-222, 241)
To set the number of reduce task write the following line of code in driver program that is ***run*** function.
job.setNumReduceTasks(?);

You can specify your partitioner let's call it JobPartitioner in run function as follows.

job.setPartitionerClass(JobPartitioner.class);

In addition to this you need to specify your Partitioner class and getPartition function in it.

```
public static class JobPartitioner extends Partitioner<IntWritable, FloatWritable > {
    @Override
    public int getPartition(IntWritable key, FloatWritable value, int numPartitions) {
        // you code goes here
        // return the appropriate value
    }
}
```

NOTE: No comparator is needed in Question 2 for now. Only provide Mapper, Reducer, Combiner and **PARTITIONER**