06 – Yash Bhatia Big Data Assignment - 6

Q.

Create 2 or 3 input files on your own , in which the data is present in different format. Write a program to process the these files using different map class and perform any one aggerate function like sum, max, min etc. on it.

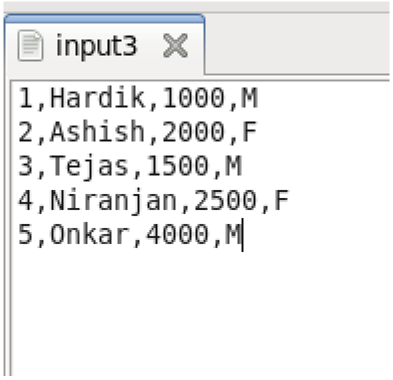
Code:

```
import java.io.IOException;

import org.apache.hadoop.conf.Configuration;

import org.apache.hadoop.fs.Path;

import org.apache.hadoop.io.IntWritable;

import org.apache.hadoop.io.LongWritable;

import org.apache.hadoop.io.Text;

import org.apache.hadoop.mapreduce.Job;

import org.apache.hadoop.mapreduce.Mapper;

import org.apache.hadoop.mapreduce.Reducer;

import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;

import org.apache.hadoop.mapreduce.lib.input.MultipleInputs;

import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;

import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;

import org.apache.hadoop.util.GenericOptionsParser;

import org.apache.commons.cli.Options;

//include external archive - hadoop-common-0.22.0.jar and commons-cli-2.0.jar


public class MultiFile

{

  public static class Map1 extends Mapper<LongWritable,Text,Text,IntWritable>
```

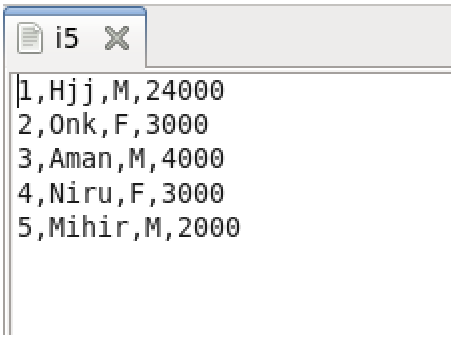
```
 {
     public void map(LongWritable key, Text value, Context con) throws IOException,
InterruptedException
     {
                    String line = value.toString();
                    String[] line1=line.split(",");
                    String gender=line1[3];
                    Text outputKey = new Text(gender);
                    int salary=Integer.parseInt(line1[2]);
                    IntWritable outputValue = new IntWritable(salary);
                    con.write(outputKey, outputValue);
     }
 }
 public static class Map2 extends Mapper<LongWritable,Text,Text,IntWritable>
 {
     public void map(LongWritable key, Text value, Context con) throws IOException,
InterruptedException
     {
                    String line = value.toString();
                    String[] line1=line.split(",");
                    String gender=line1[2];
                    Text outputKey = new Text(gender);
                    int salary=Integer.parseInt(line1[3]);
                    IntWritable outputValue = new IntWritable(salary);
                    con.write(outputKey, outputValue);
     }
```

```
}

public static class Red extends Reducer<Text,IntWritable,Text,IntWritable>

{

        public void reduce(Text gender, Iterable<IntWritable> total_sal, Context con)

         throws IOException , InterruptedException

         {


                        int sum = 0;

                        for(IntWritable value : total_sal)

                        {

                                sum += value.get();

                        }

                        con.write(gender, new IntWritable(sum));


         }

 }

public static void main(String[] args) throws Exception

{

        Configuration c=new Configuration();

        GenericOptionsParser parser= new GenericOptionsParser(c,args);

        String[] files= parser.getRemainingArgs();

        Path p1=new Path(files[0]);

        Path p2=new Path(files[1]);

        Path p3=new Path(files[2]);

        Job j = new Job(c,"multiple");
        j.setJarByClass(MultiFile.class);

        j.setMapperClass(Map1.class);

        j.setMapperClass(Map2.class);

        j.setReducerClass(Red.class);

        j.setOutputKeyClass(Text.class);

        j.setOutputValueClass(IntWritable.class);

        MultipleInputs.addInputPath(j, p1, TextInputFormat.class, Map1.class);

        MultipleInputs.addInputPath(j,p2, TextInputFormat.class, Map2.class);

    FileOutputFormat.setOutputPath(j, p3);

    System.exit(j.waitForCompletion(true) ? 0:1);

       }}
```

Input files:

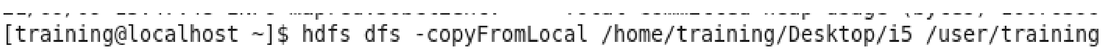


```
1,Hardik,1000,M
2,Ashish,2000,F
3,Tejas,1500,M
4,Niranjan,2500,F
5,Onkar,4000,M
```

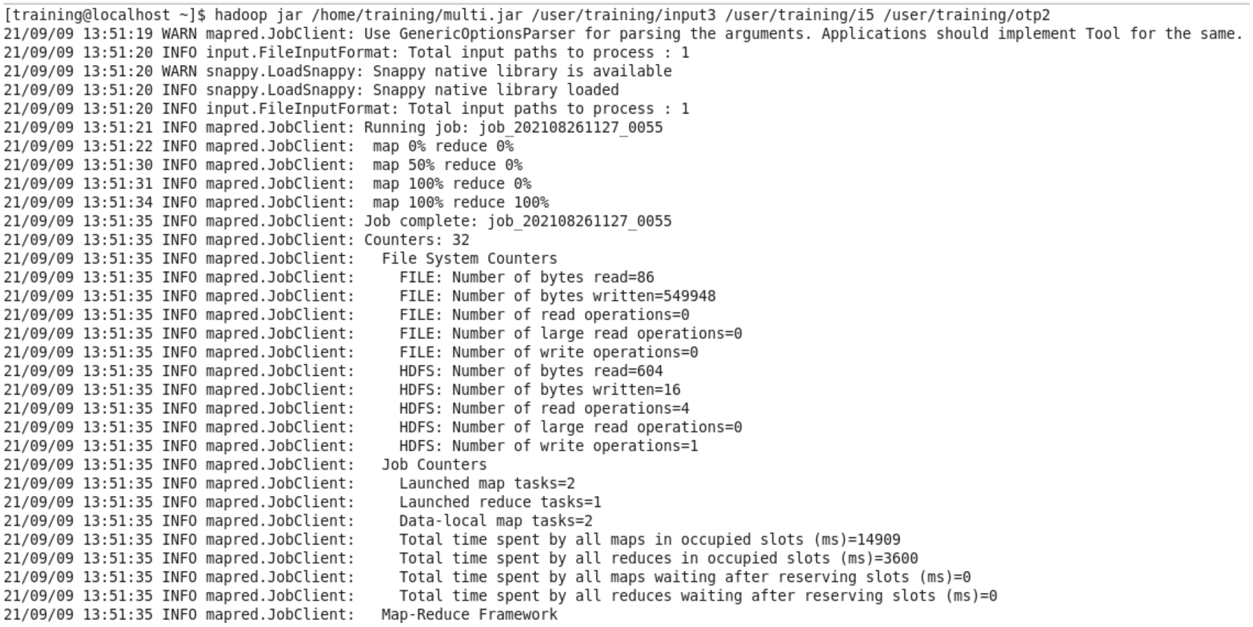


```
1,Hjj,M,24000
2,Onk,F,3000
3,Aman,M,4000
4,Niru,F,3000
5,Mihir,M,2000
```

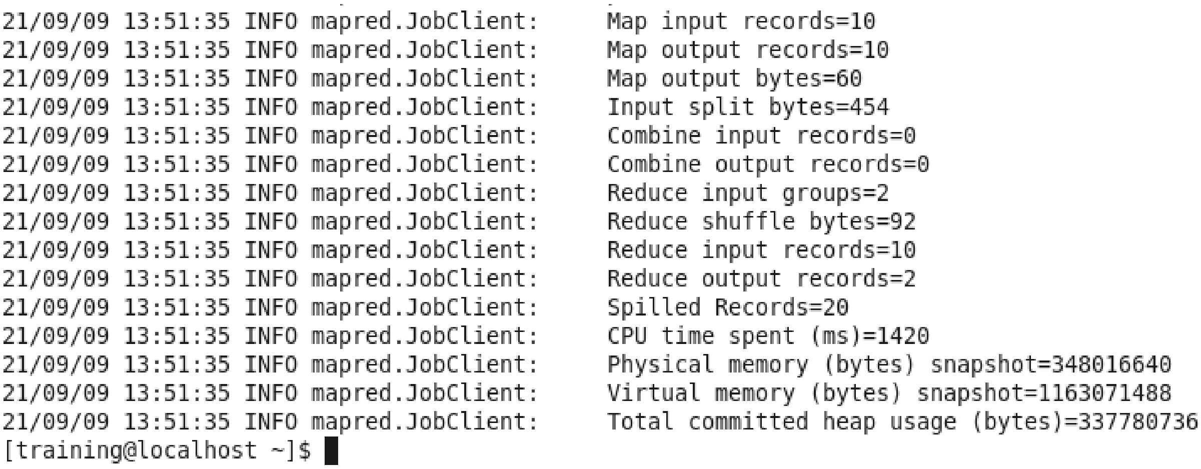
Command Line Screenshots:



```
[training@localhost ~]$ hdfs dfs -copyFromLocal /home/training/Desktop/i5 /user/training
```

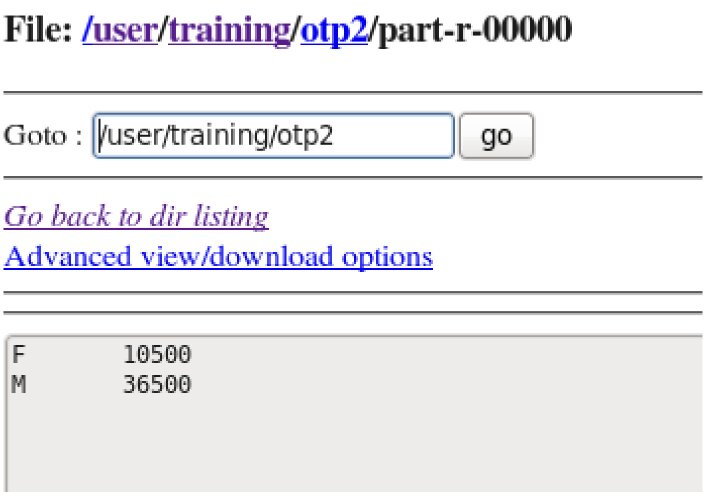


```
[training@localhost ~]$ hadoop jar /home/training/multi.jar /user/training/input3 /user/training/i5 /user/training/otp2
21/09/09 13:51:19 WARN mapred.JobClient: Use GenericOptionsParser for parsing the arguments. Applications should implement Tool for the same.
21/09/09 13:51:20 INFO input.FileInputFormat: Total input paths to process : 1
21/09/09 13:51:20 WARN snappy.LoadSnappy: Snappy native library is available
21/09/09 13:51:20 INFO snappy.LoadSnappy: Snappy native library loaded
21/09/09 13:51:20 INFO input.FileInputFormat: Total input paths to process : 1
21/09/09 13:51:21 INFO mapred.JobClient: Running job: job_202108261127_0055
21/09/09 13:51:22 INFO mapred.JobClient:  map 0% reduce 0%
21/09/09 13:51:30 INFO mapred.JobClient:  map 50% reduce 0%
21/09/09 13:51:31 INFO mapred.JobClient:  map 100% reduce 0%
21/09/09 13:51:34 INFO mapred.JobClient:  map 100% reduce 100%
21/09/09 13:51:35 INFO mapred.JobClient: Job complete: job_202108261127_0055
21/09/09 13:51:35 INFO mapred.JobClient: Counters: 32
21/09/09 13:51:35 INFO mapred.JobClient:   File System Counters
21/09/09 13:51:35 INFO mapred.JobClient:     FILE: Number of bytes read=86
21/09/09 13:51:35 INFO mapred.JobClient:     FILE: Number of bytes written=549948
21/09/09 13:51:35 INFO mapred.JobClient:     FILE: Number of read operations=0
21/09/09 13:51:35 INFO mapred.JobClient:     FILE: Number of large read operations=0
21/09/09 13:51:35 INFO mapred.JobClient:     FILE: Number of write operations=0
21/09/09 13:51:35 INFO mapred.JobClient:     HDFS: Number of bytes read=604
21/09/09 13:51:35 INFO mapred.JobClient:     HDFS: Number of bytes written=16
21/09/09 13:51:35 INFO mapred.JobClient:     HDFS: Number of read operations=4
21/09/09 13:51:35 INFO mapred.JobClient:     HDFS: Number of large read operations=0
21/09/09 13:51:35 INFO mapred.JobClient:     HDFS: Number of write operations=1
21/09/09 13:51:35 INFO mapred.JobClient:   Job Counters
21/09/09 13:51:35 INFO mapred.JobClient:     Launched map tasks=2
21/09/09 13:51:35 INFO mapred.JobClient:     Launched reduce tasks=1
21/09/09 13:51:35 INFO mapred.JobClient:     Data-local map tasks=2
21/09/09 13:51:35 INFO mapred.JobClient:     Total time spent by all maps in occupied slots (ms)=14909
21/09/09 13:51:35 INFO mapred.JobClient:     Total time spent by all reduces in occupied slots (ms)=3600
21/09/09 13:51:35 INFO mapred.JobClient:     Total time spent by all maps waiting after reserving slots (ms)=0
21/09/09 13:51:35 INFO mapred.JobClient:     Total time spent by all reduces waiting after reserving slots (ms)=0
21/09/09 13:51:35 INFO mapred.JobClient:   Map-Reduce Framework
```



```
21/09/09 13:51:35 INFO mapred.JobClient:     Map input records=10
21/09/09 13:51:35 INFO mapred.JobClient:     Map output records=10
21/09/09 13:51:35 INFO mapred.JobClient:     Map output bytes=60
21/09/09 13:51:35 INFO mapred.JobClient:     Input split bytes=454
21/09/09 13:51:35 INFO mapred.JobClient:     Combine input records=0
21/09/09 13:51:35 INFO mapred.JobClient:     Combine output records=0
21/09/09 13:51:35 INFO mapred.JobClient:     Reduce input groups=2
21/09/09 13:51:35 INFO mapred.JobClient:     Reduce shuffle bytes=92
21/09/09 13:51:35 INFO mapred.JobClient:     Reduce input records=10
21/09/09 13:51:35 INFO mapred.JobClient:     Reduce output records=2
21/09/09 13:51:35 INFO mapred.JobClient:     Spilled Records=20
21/09/09 13:51:35 INFO mapred.JobClient:     CPU time spent (ms)=1420
21/09/09 13:51:35 INFO mapred.JobClient:     Physical memory (bytes) snapshot=348016640
21/09/09 13:51:35 INFO mapred.JobClient:     Virtual memory (bytes) snapshot=1163071488
21/09/09 13:51:35 INFO mapred.JobClient:     Total committed heap usage (bytes)=337780736
[training@localhost ~]$ █
```

Output:



File: /user/training/otp2/part-r-00000

Goto : /user/training/otp2

Go back to dir listing

Advanced view/download options

```
F       10500
M       36500
```