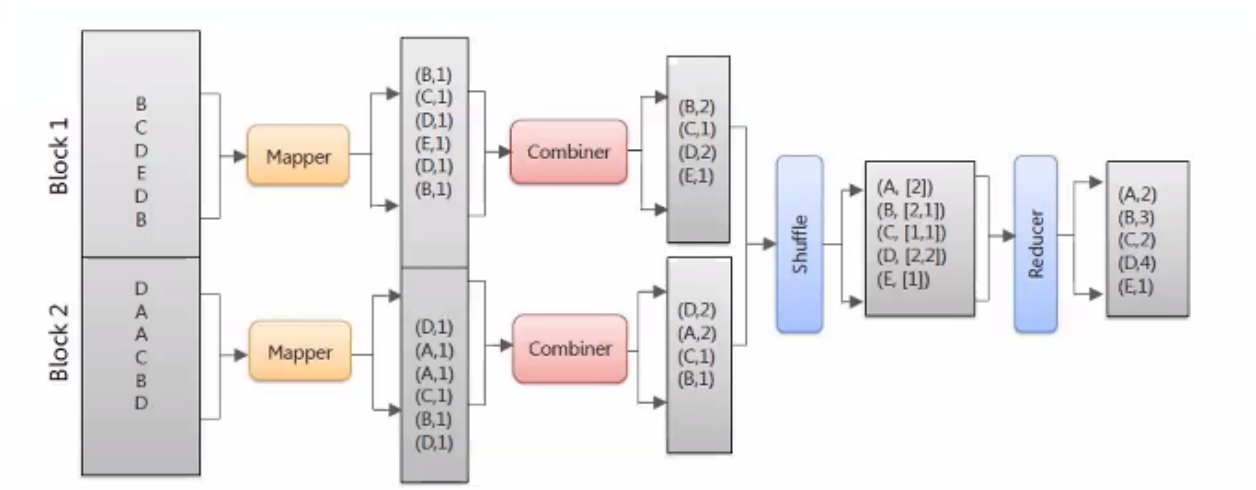
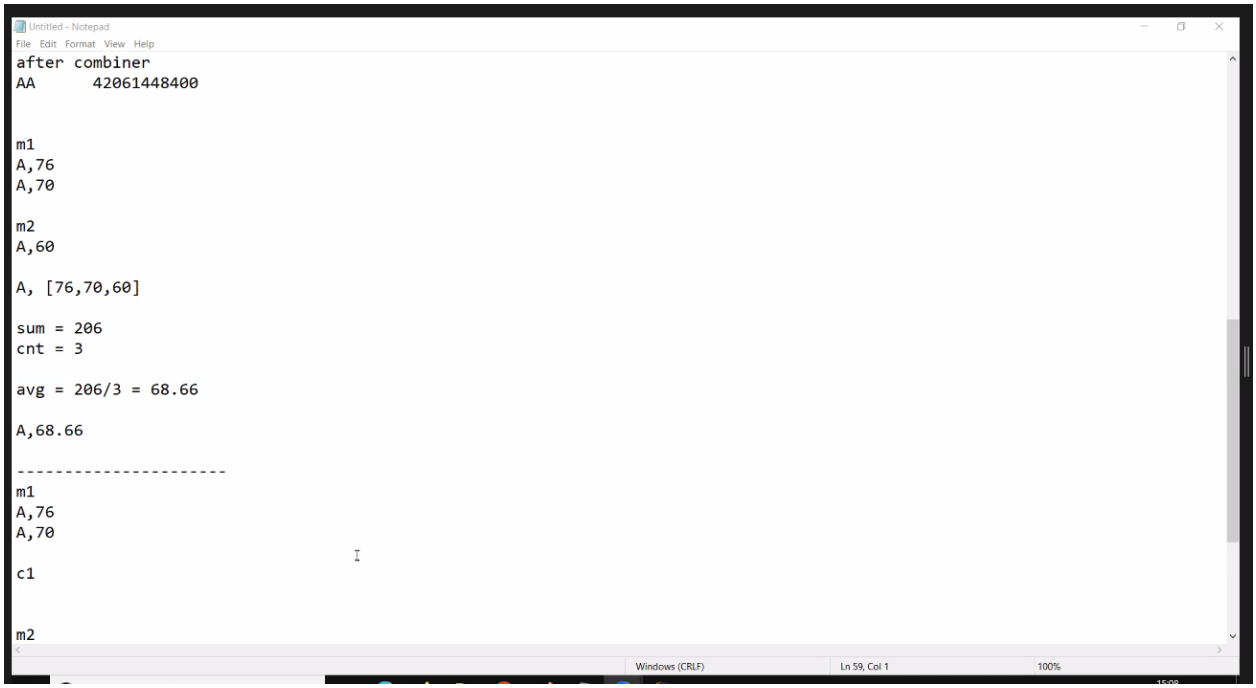
→ Combiner

a. Called local reducer
b. Runs on same node where mapper is running
c. Combines data





→

```
import java.io.*;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.Mapper;
import org.apache.hadoop.mapreduce.Reducer;
import org.apache.hadoop.conf.*;
import org.apache.hadoop.fs.*;
import org.apache.hadoop.mapreduce.lib.input.*;
import org.apache.hadoop.mapreduce.lib.output.*;
public class StockVolumeWithCombiner {

        public static class MapClass extends Mapper<LongWritable,Text,Text,LongWritable>
          {
            public void map(LongWritable key, Text value, Context context)
            {
              try{
                String[] str = value.toString().split(",");
                long vol = Long.parseLong(str[7]);
                context.write(new Text(str[1]),new LongWritable(vol));
              }
              catch(Exception e)
```

```
                {
                    System.out.println(e.getMessage());
                }
            }
        }

        public static class ReduceClass extends Reducer<Text,LongWritable,Text,LongWritable>
        {
                private LongWritable result = new LongWritable();

                public void reduce(Text key, Iterable<LongWritable> values,Context context) throws
IOException, InterruptedException {
                    long sum = 0;

                      for (LongWritable val : values)
                      {
                          sum += val.get();
                      }

                    result.set(sum);
                    context.write(key, result);
                    //context.write(key, new LongWritable(sum));

                }
        }
        public static void main(String[] args) throws Exception {
                Configuration conf = new Configuration();
                //conf.set("name", "value")
                //conf.set("mapreduce.input.fileinputformat.split.minsize", "134217728");
                Job job = Job.getInstance(conf, "Volume Count");
                job.setJarByClass(StockVolume.class);
                job.setMapperClass(MapClass.class);
                //job.setCombinerClass(ReduceClass.class);
                job.setCombinerClass(ReduceClass.class); //enables combiner class to run at mapper
                job.setReducerClass(ReduceClass.class);
//              job.setNumReduceTasks(1); // means it will run one reducer only
                job.setNumReduceTasks(0); // means no reducer will run
                job.setOutputKeyClass(Text.class);
                job.setOutputValueClass(LongWritable.class);
                FileInputFormat.addInputPath(job, new Path(args[0]));
                FileOutputFormat.setOutputPath(job, new Path(args[1]));
                System.exit(job.waitForCompletion(true) ? 0 : 1);
            }
}
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ jar tvf myjar.jar
   25 Sat May 20 15:05:20 UTC 2023 META-INF/MANIFEST.MF
```

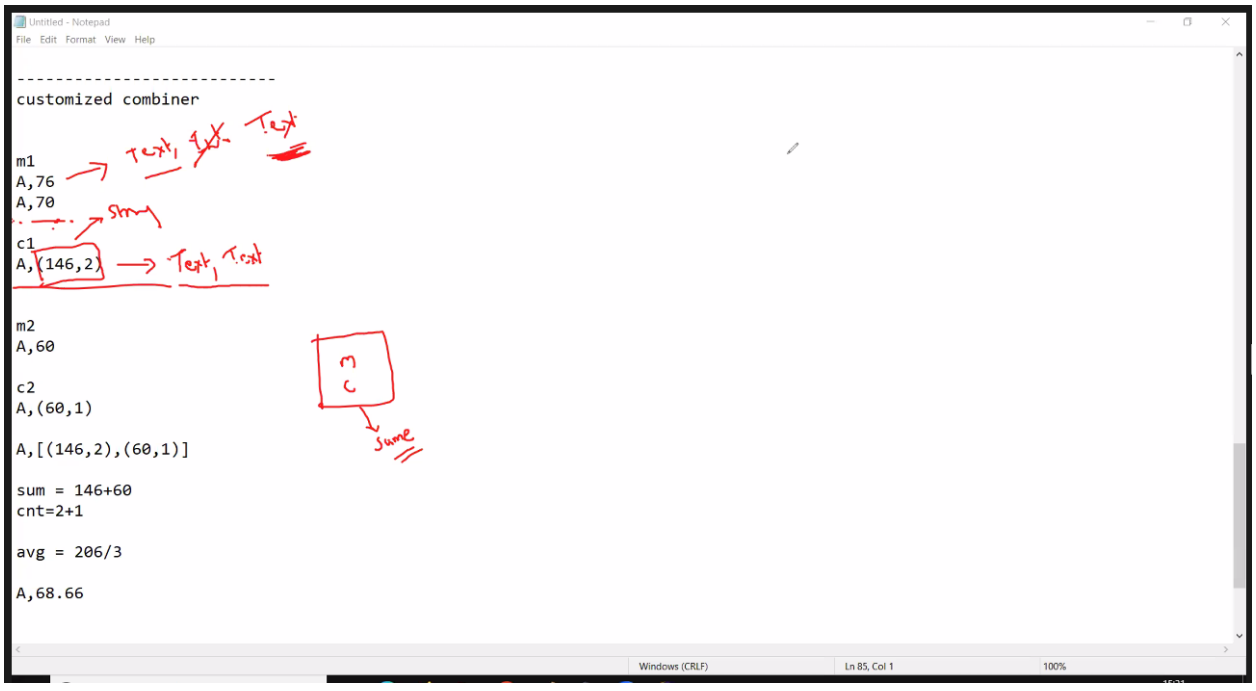
```
 2459 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$MapClass.class
 2392 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$ReduceClass.class
 1722 Fri May 19 16:03:30 UTC 2023 AllTimeHigh.class
 2475 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$MapClass.class
 2454 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$ReduceClass.class
 1732 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice.class
 2337 Fri May 19 17:41:44 UTC 2023 WordCount$IntSumReducer.class
 2461 Fri May 19 17:41:44 UTC 2023 WordCount$TokenizerMapper.class
 1790 Fri May 19 17:41:44 UTC 2023 WordCount.class
 2454 Fri May 19 15:53:50 UTC 2023 AllTimeLow$MapClass.class
 2388 Fri May 19 15:53:50 UTC 2023 AllTimeLow$ReduceClass.class
 1734 Fri May 19 15:53:50 UTC 2023 AllTimeLow.class
 2408 Thu May 18 17:48:56 UTC 2023 StockVolume$MapClass.class
 2349 Thu May 18 17:48:56 UTC 2023 StockVolume$ReduceClass.class
 1697 Thu May 18 17:48:56 UTC 2023 StockVolume.class
 2456 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$MapClass.class
 2397 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$ReduceClass.class
 1813 Sat May 20 15:04:24 UTC 2023 StockVolumeWithCombiner.class
```

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop jar myjar.jar
StockVolumeWithCombiner training/NYSE.csv training/out7
WARNING: Use "yarn jar" to launch YARN applications.
23/05/20 09:39:36 INFO client.RMProxy: Connecting to ResourceManager
at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/20 09:39:37 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the Tool
interface and execute your application with T
oolRunner to remedy this.
23/05/20 09:39:37 INFO mapreduce.JobResourceUploader: Disabling
Erasure Coding for path:
/user/bigdatalab456422/.staging/job_1684298513961_0758
23/05/20 09:39:37 INFO input.FileInputFormat: Total input files to
process : 1
23/05/20 09:39:37 INFO mapreduce.JobSubmitter: number of splits:1
23/05/20 09:39:37 INFO Configuration.deprecation:
yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enable
d
23/05/20 09:39:37 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1684298513961_0758
```

```
23/05/20 09:39:37 INFO mapreduce.JobSubmitter: Executing with tokens:
[]
23/05/20 09:39:37 INFO conf.Configuration: resource-types.xml not
found
23/05/20 09:39:37 INFO resource.ResourceUtils: Unable to find
'resource-types.xml'.
23/05/20 09:39:37 INFO impl.YarnClientImpl: Submitted application
application_1684298513961_0758
23/05/20 09:39:37 INFO mapreduce.Job: The url to track the job:
http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/applicati
on_1684298513961_0758/
23/05/20 09:39:37 INFO mapreduce.Job: Running job:
job_1684298513961_0758
23/05/20 09:39:47 INFO mapreduce.Job: Job job_1684298513961_0758
running in uber mode : false
23/05/20 09:39:47 INFO mapreduce.Job:  map 0% reduce 0%
23/05/20 09:39:53 INFO mapreduce.Job:  map 100% reduce 0%
23/05/20 09:39:53 INFO mapreduce.Job: Job job_1684298513961_0758
completed successfully
23/05/20 09:39:53 INFO mapreduce.Job: Counters: 33
       File System Counters
               FILE: Number of bytes read=0
               FILE: Number of bytes written=222643
               FILE: Number of read operations=0
               FILE: Number of large read operations=0
               FILE: Number of write operations=0
               HDFS: Number of bytes read=40990986
               HDFS: Number of bytes written=7842509
               HDFS: Number of read operations=7
               HDFS: Number of large read operations=0
               HDFS: Number of write operations=2
               HDFS: Number of bytes read erasure-coded=0
       Job Counters
               Launched map tasks=1
               Data-local map tasks=1
               Total time spent by all maps in occupied slots
(ms)=4110
               Total time spent by all reduces in occupied slots
(ms)=0
               Total time spent by all map tasks (ms)=4110
               Total vcore-milliseconds taken by all map tasks=4110
               Total megabyte-milliseconds taken by all map
tasks=4208640
       Map-Reduce Framework
               Map input records=735026
```

```
		Map output records=735026
		Input split bytes=124
		Spilled Records=0
		Failed Shuffles=0
		Merged Map outputs=0
		GC time elapsed (ms)=68
		CPU time spent (ms)=3180
		Physical memory (bytes) snapshot=370642944
		Virtual memory (bytes) snapshot=2587979776
		Total committed heap usage (bytes)=468713472
		Peak Map Physical memory (bytes)=370642944
		Peak Map Virtual memory (bytes)=2587979776
	File Input Format Counters
		Bytes Read=40990862
	File Output Format Counters
		Bytes Written=7842509
```

→ customized combiner



→ customized combiner: find out average score for students of each city (check if Pune city has average score 68.6666…)

#

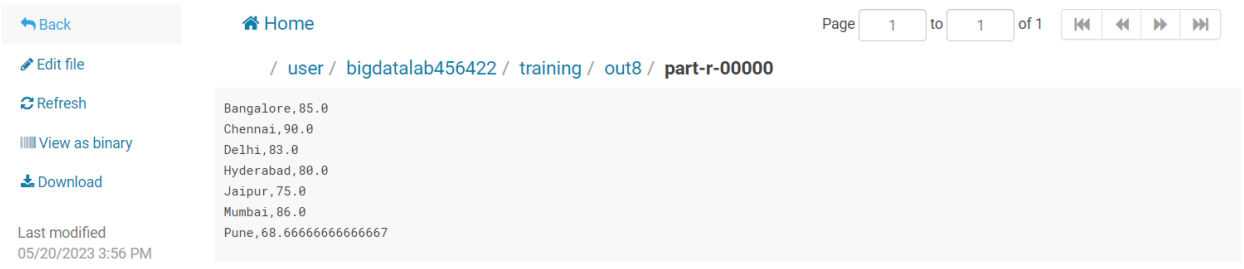
```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -mkdir student
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop fs -put student1 student2
student
```

#

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop jar myjar.jar CityAvg2
student training/out8
WARNING: Use "yarn jar" to launch YARN applications.
23/05/20 10:25:44 INFO client.RMProxy: Connecting to ResourceManager
at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/20 10:25:45 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the Tool
interface and execute your application with T
oolRunner to remedy this.
23/05/20 10:25:45 INFO mapreduce.JobResourceUploader: Disabling
Erasure Coding for path:
/user/bigdatalab456422/.staging/job_1684298513961_0784
23/05/20 10:25:45 INFO input.FileInputFormat: Total input files to
process : 2
23/05/20 10:25:46 INFO mapreduce.JobSubmitter: number of splits:2
23/05/20 10:25:46 INFO Configuration.deprecation:
yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enable
d
23/05/20 10:25:46 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1684298513961_0784
23/05/20 10:25:46 INFO mapreduce.JobSubmitter: Executing with tokens:
[]
23/05/20 10:25:46 INFO conf.Configuration: resource-types.xml not
found
23/05/20 10:25:46 INFO resource.ResourceUtils: Unable to find
'resource-types.xml'.
23/05/20 10:25:46 INFO impl.YarnClientImpl: Submitted application
application_1684298513961_0784
23/05/20 10:25:46 INFO mapreduce.Job: The url to track the job:
http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/applicati
on_1684298513961_0784/
23/05/20 10:25:46 INFO mapreduce.Job: Running job:
job_1684298513961_0784
23/05/20 10:26:09 INFO mapreduce.Job: Job job_1684298513961_0784
running in uber mode : false
23/05/20 10:26:09 INFO mapreduce.Job:  map 0% reduce 0%
```

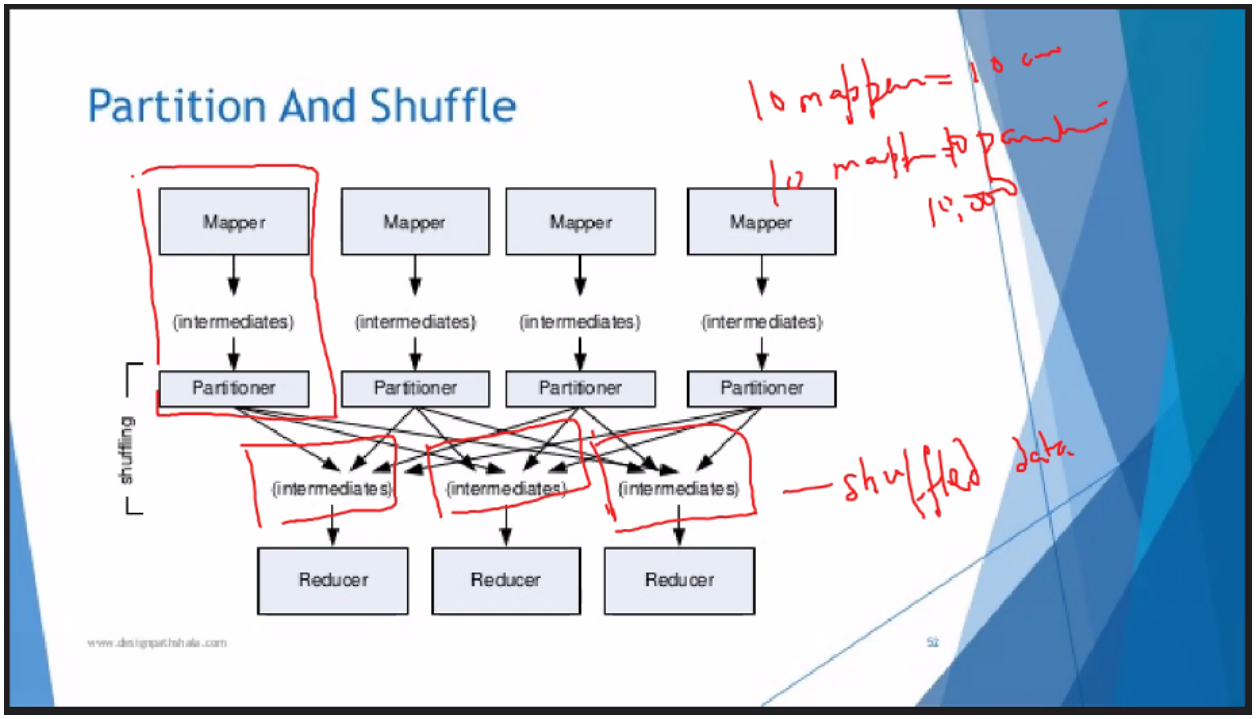
```
23/05/20 10:26:17 INFO mapreduce.Job:  map 100% reduce 0%
23/05/20 10:26:31 INFO mapreduce.Job:  map 100% reduce 100%
23/05/20 10:26:33 INFO mapreduce.Job: Job job_1684298513961_0784
completed successfully
23/05/20 10:26:34 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=148
                FILE: Number of bytes written=670328
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=493
                HDFS: Number of bytes written=101
                HDFS: Number of read operations=11
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Data-local map tasks=2
                Total time spent by all maps in occupied slots
(ms)=10570
                Total time spent by all reduces in occupied slots
(ms)=12278
                Total time spent by all map tasks (ms)=10570
                Total time spent by all reduce tasks (ms)=12278
                Total vcore-milliseconds taken by all map tasks=10570
                Total vcore-milliseconds taken by all reduce
tasks=12278
                Total megabyte-milliseconds taken by all map
tasks=10823680
                Total megabyte-milliseconds taken by all reduce
tasks=12572672
        Map-Reduce Framework
                Map input records=12
                Map output records=12
                Map output bytes=122
                Map output materialized bytes=179
                Input split bytes=246
                Combine input records=12
                Combine output records=10
                Reduce input groups=7
                Reduce shuffle bytes=179
                Reduce input records=10
```

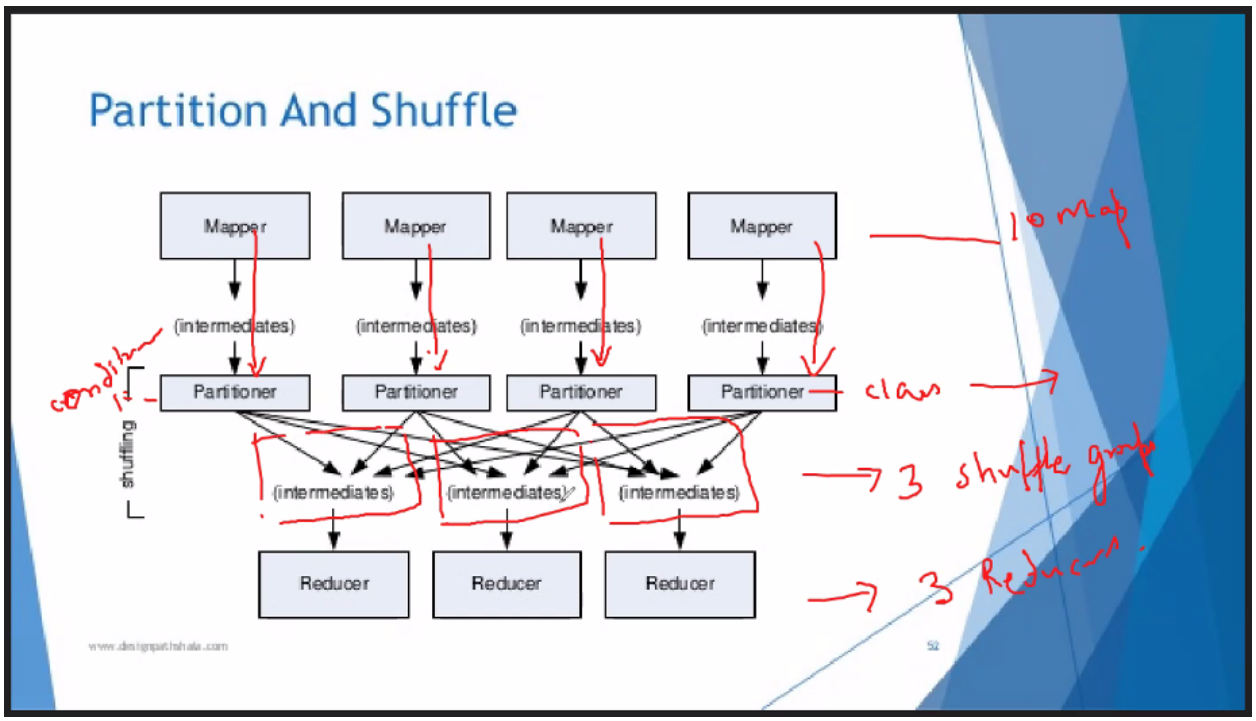
```
        Reduce output records=7
        Spilled Records=20
        Shuffled Maps =2
        Failed Shuffles=0
        Merged Map outputs=2
        GC time elapsed (ms)=571
        CPU time spent (ms)=3370
        Physical memory (bytes) snapshot=1306103808
        Virtual memory (bytes) snapshot=7778222080
        Total committed heap usage (bytes)=1710227456
        Peak Map Physical memory (bytes)=526311424
        Peak Map Virtual memory (bytes)=2590330880
        Peak Reduce Physical memory (bytes)=256356352
        Peak Reduce Virtual memory (bytes)=2599239680
Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
File Input Format Counters
        Bytes Read=247
File Output Format Counters
        Bytes Written=101
```

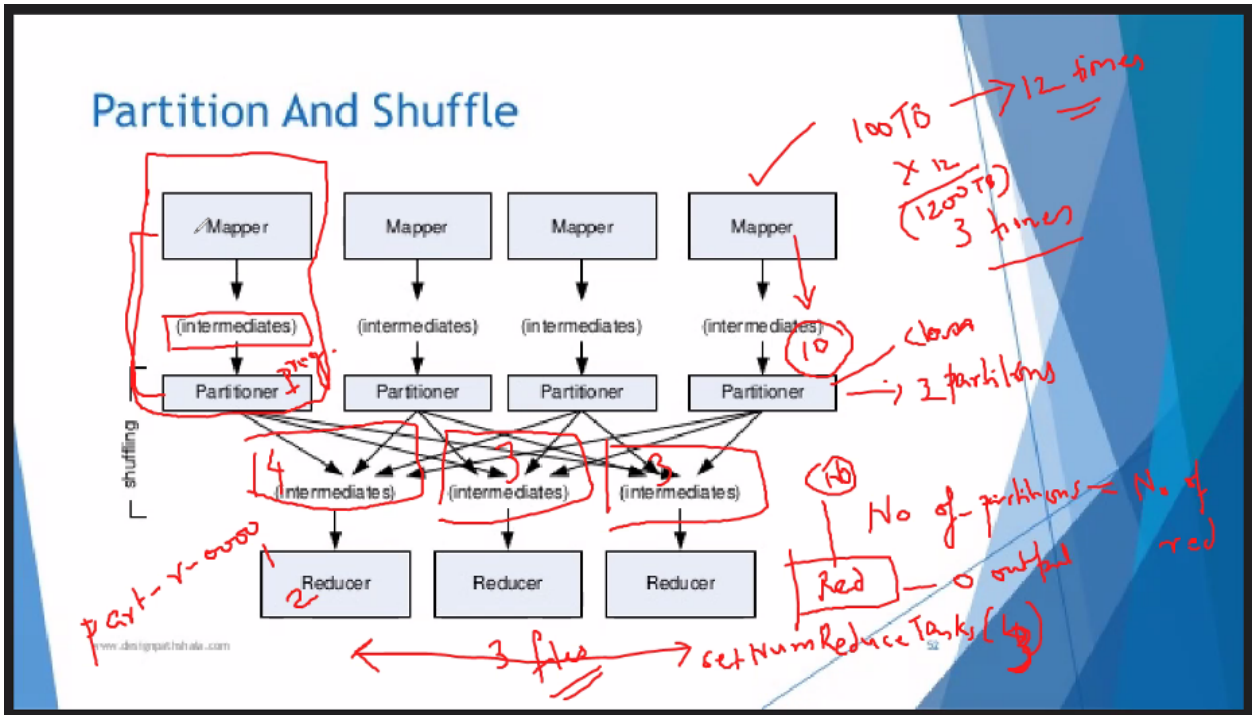


→ Partitioner

a. Called local reducer
b. Runs on same node where mapper is running
c.







emp.csv - Notepad

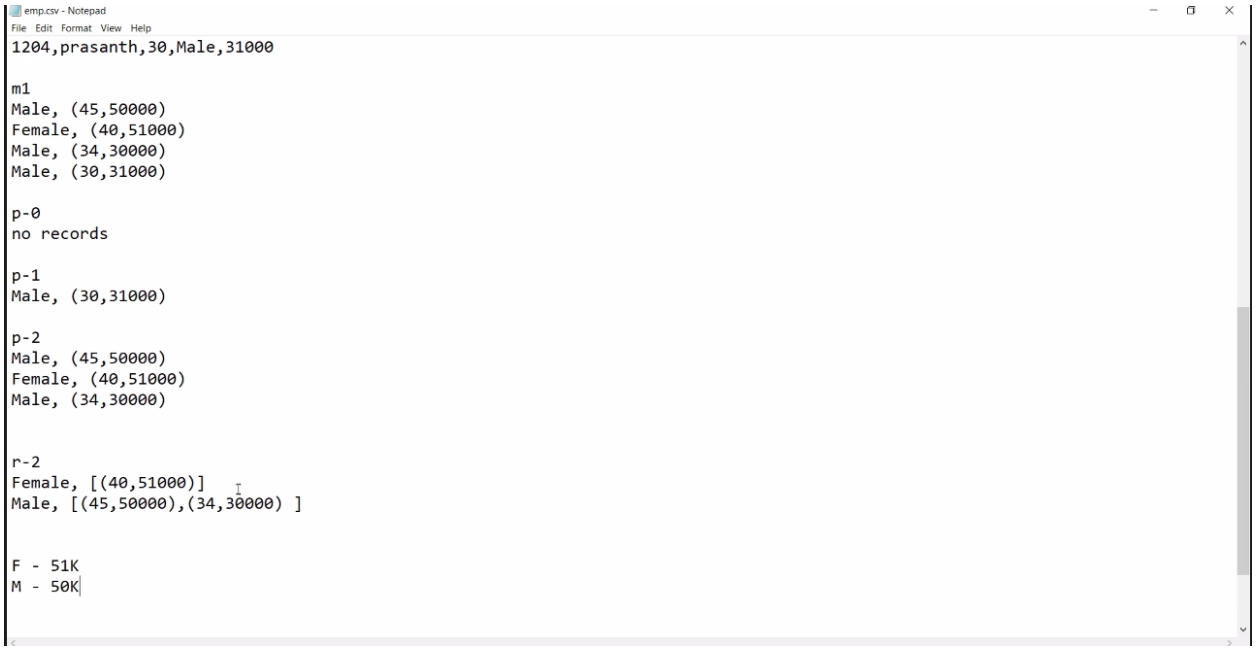
File Edit Format View Help

```
1201,gopal,45,Male,50000
1202,manisha,40,Female,51000
1203,khaleel,34,Male,30000
1204,prasanth,30,Male,31000
1205,kiran,20,Male,40000
1206,laxmi,25,Female,35000
1207,bhavya,20,Female,15000
1208,reshma,19,Female,14000
1209,kranthi,22,Male,22000
1210,Satish,24,Male,25000
1211,Krishna,25,Male,26000
1212,Arshad,28,Male,20000
1213,lavanya,18,Female,8000

prob statement : find max salary paid to each gender in 3 diff age groups

1) age<=20
2) age > and <=30
3) age > 30


mapper code
gender, (age,sal)
```



→ find max salary paid to each gender in 3 diff age groups

1) age<=20
2) age > 20 and age <= 30
3) age > 30

#

```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop  fs -put emp.csv training
```

#

```
[bigdatalab456422@ip-10-1-1-204 ~]$ jar tvf myjar.jar
   25 Sat May 20 17:36:10 UTC 2023 META-INF/MANIFEST.MF
 2459 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$MapClass.class
 2392 Fri May 19 16:03:30 UTC 2023 AllTimeHigh$ReduceClass.class
 1722 Fri May 19 16:03:30 UTC 2023 AllTimeHigh.class
 2475 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$MapClass.class
 2454 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice$ReduceClass.class
 1732 Fri May 19 16:53:46 UTC 2023 AvgClosingPrice.class
 2337 Fri May 19 17:41:44 UTC 2023 WordCount$IntSumReducer.class
 2461 Fri May 19 17:41:44 UTC 2023 WordCount$TokenizerMapper.class
 1790 Fri May 19 17:41:44 UTC 2023 WordCount.class
 2454 Fri May 19 15:53:50 UTC 2023 AllTimeLow$MapClass.class
 2388 Fri May 19 15:53:50 UTC 2023 AllTimeLow$ReduceClass.class
 1734 Fri May 19 15:53:50 UTC 2023 AllTimeLow.class
 1242 Sat May 20 17:28:50 UTC 2023
MyPartitioner$CaderPartitioner.class
 2365 Sat May 20 17:28:50 UTC 2023 MyPartitioner$MapClass.class
 2905 Sat May 20 17:28:50 UTC 2023 MyPartitioner$ReduceClass.class
 2632 Sat May 20 17:28:50 UTC 2023 MyPartitioner.class
 2408 Thu May 18 17:48:56 UTC 2023 StockVolume$MapClass.class
 2349 Thu May 18 17:48:56 UTC 2023 StockVolume$ReduceClass.class
 1697 Thu May 18 17:48:56 UTC 2023 StockVolume.class
 2648 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityCombineClass.class
 2269 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityMapClass.class
 2639 Sat May 20 15:42:48 UTC 2023 CityAvg2$CityReduceClass.class
 2034 Sat May 20 15:42:48 UTC 2023 CityAvg2.class
 2456 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$MapClass.class
 2397 Sat May 20 15:04:24 UTC 2023
StockVolumeWithCombiner$ReduceClass.class
 1813 Sat May 20 15:04:24 UTC 2023 StockVolumeWithCombiner.class
```

#

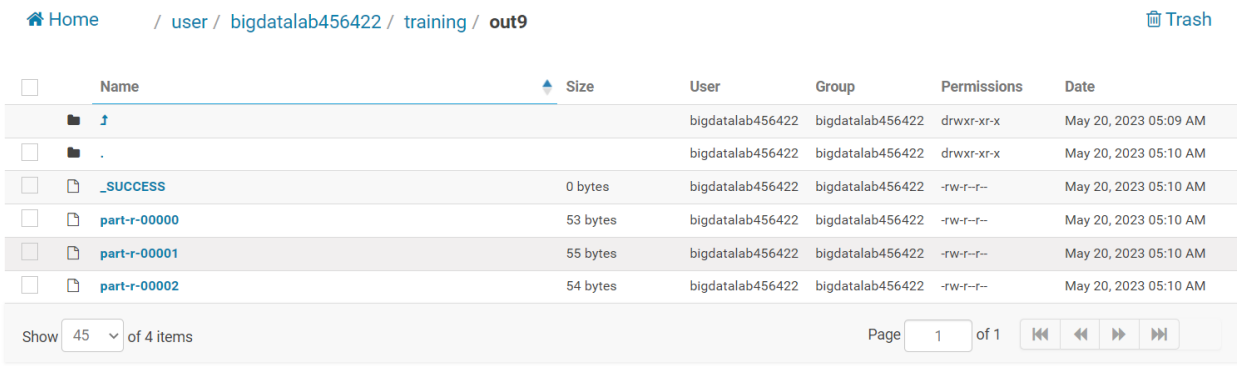
```
[bigdatalab456422@ip-10-1-1-204 ~]$ hadoop jar myjar.jar
MyPartitioner training/emp.csv training/out9
WARNING: Use "yarn jar" to launch YARN applications.
23/05/20 12:09:40 INFO client.RMProxy: Connecting to ResourceManager
at ip-10-1-1-204.ap-south-1.compute.internal/10.1.1.204:8032
23/05/20 12:09:40 WARN mapreduce.JobResourceUploader: Hadoop
command-line option parsing not performed. Implement the Tool
interface and execute your application with T
```

```
oolRunner to remedy this.
23/05/20 12:09:40 INFO mapreduce.JobResourceUploader: Disabling
Erasure Coding for path:
/user/bigdatalab456422/.staging/job_1684298513961_0866
23/05/20 12:09:41 INFO input.FileInputFormat: Total input files to
process : 1
23/05/20 12:09:41 INFO mapreduce.JobSubmitter: number of splits:1
23/05/20 12:09:41 INFO Configuration.deprecation:
yarn.resourcemanager.system-metrics-publisher.enabled is deprecated.
Instead, use yarn.system-metrics-publisher.enable
d
23/05/20 12:09:41 INFO mapreduce.JobSubmitter: Submitting tokens for
job: job_1684298513961_0866
23/05/20 12:09:41 INFO mapreduce.JobSubmitter: Executing with tokens:
[]
23/05/20 12:09:41 INFO conf.Configuration: resource-types.xml not
found
23/05/20 12:09:41 INFO resource.ResourceUtils: Unable to find
'resource-types.xml'.
23/05/20 12:09:42 INFO impl.YarnClientImpl: Submitted application
application_1684298513961_0866
23/05/20 12:09:42 INFO mapreduce.Job: The url to track the job:
http://ip-10-1-1-204.ap-south-1.compute.internal:6066/proxy/applicati
on_1684298513961_0866/
23/05/20 12:09:42 INFO mapreduce.Job: Running job:
job_1684298513961_0866
23/05/20 12:09:57 INFO mapreduce.Job: Job job_1684298513961_0866
running in uber mode : false
23/05/20 12:09:57 INFO mapreduce.Job:  map 0% reduce 0%
23/05/20 12:10:04 INFO mapreduce.Job:  map 100% reduce 0%
23/05/20 12:10:13 INFO mapreduce.Job:  map 100% reduce 67%
23/05/20 12:10:16 INFO mapreduce.Job:  map 100% reduce 100%
23/05/20 12:10:17 INFO mapreduce.Job: Job job_1684298513961_0866
completed successfully
23/05/20 12:10:17 INFO mapreduce.Job: Counters: 54
       File System Counters
                FILE: Number of bytes read=434
                FILE: Number of bytes written=895851
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=474
                HDFS: Number of bytes written=162
                HDFS: Number of read operations=18
                HDFS: Number of large read operations=0
```

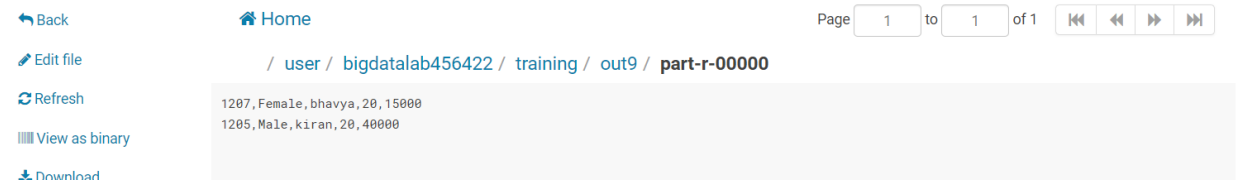
```
                HDFS: Number of write operations=6
                HDFS: Number of bytes read erasure-coded=0
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=3
                Data-local map tasks=1
                Total time spent by all maps in occupied slots
(ms)=4465
                Total time spent by all reduces in occupied slots
(ms)=24744
                Total time spent by all map tasks (ms)=4465
                Total time spent by all reduce tasks (ms)=24744
                Total vcore-milliseconds taken by all map tasks=4465
                Total vcore-milliseconds taken by all reduce
tasks=24744
                Total megabyte-milliseconds taken by all map
tasks=4572160
                Total megabyte-milliseconds taken by all reduce
tasks=25337856
        Map-Reduce Framework
                Map input records=13
                Map output records=13
                Map output bytes=426
                Map output materialized bytes=422
                Input split bytes=123
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=422
                Reduce input records=13
                Reduce output records=6
                Spilled Records=26
                Shuffled Maps =3
                Failed Shuffles=0
                Merged Map outputs=3
                GC time elapsed (ms)=697
                CPU time spent (ms)=4550
                Physical memory (bytes) snapshot=1158127616
                Virtual memory (bytes) snapshot=10370674688
                Total committed heap usage (bytes)=1611137024
                Peak Map Physical memory (bytes)=526077952
                Peak Map Virtual memory (bytes)=2587410432
                Peak Reduce Physical memory (bytes)=249159680
                Peak Reduce Virtual memory (bytes)=2601119744
        Shuffle Errors
```

```
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
        WRONG_REDUCE=0
    File Input Format Counters
        Bytes Read=351
    File Output Format Counters
        Bytes Written=162
```

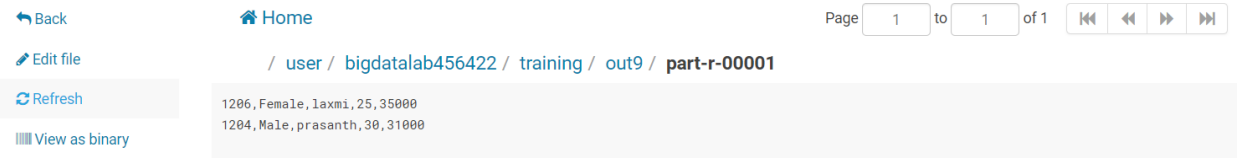
creates three folders for three outputs from respective three reducers



for age group0 [age<=20]



for age group1 [age > 20 and age <= 30]



for age group2 [age > 30]