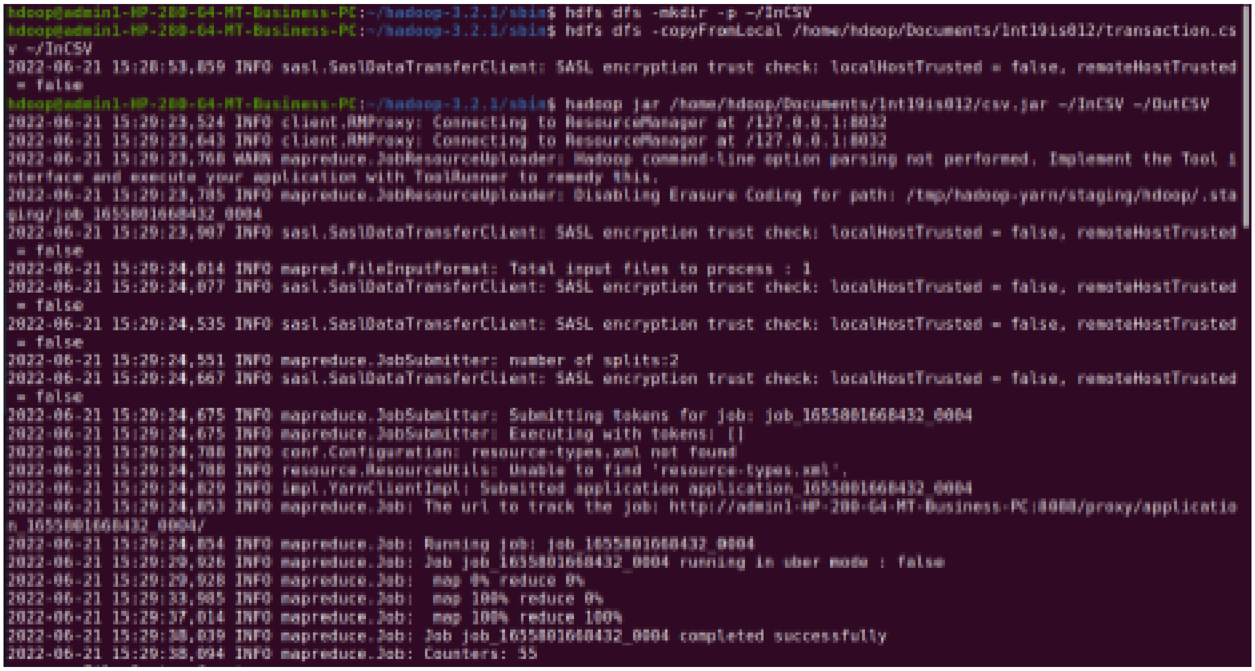
Vansh Sharma
1NT19IS181
C2 Batch

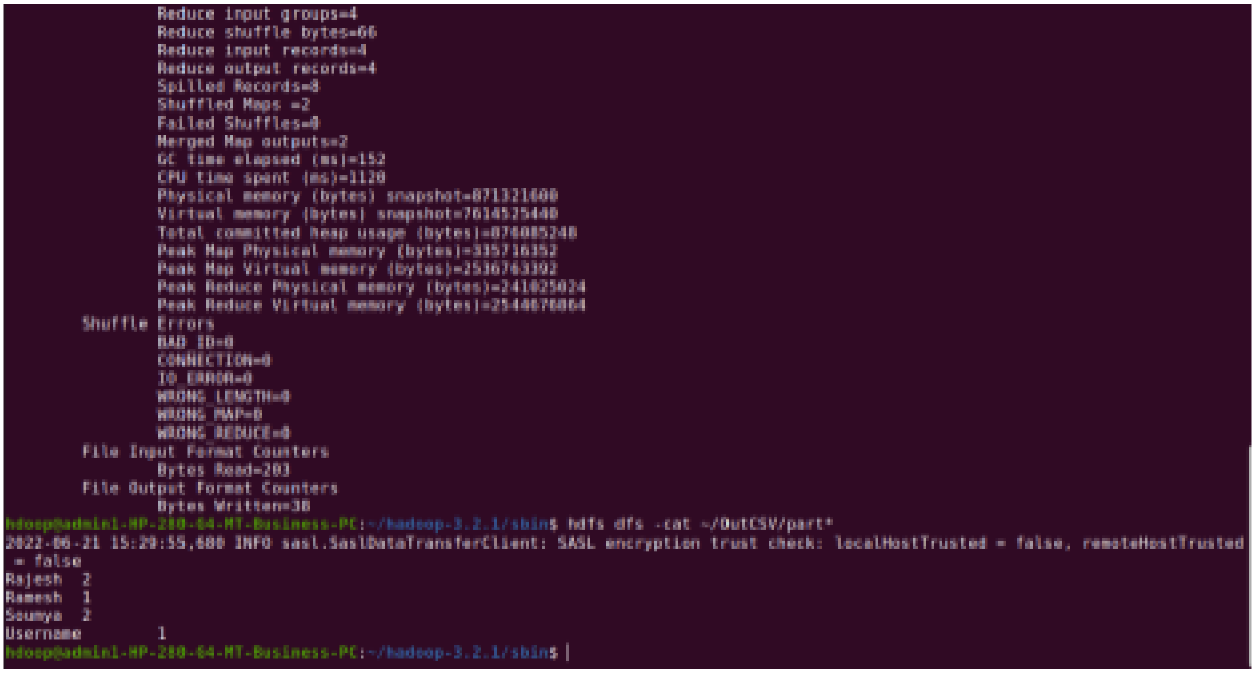
Date:16/06/2022

Hadoop CSV Map Reduce
Part A

```
package csv;
import java.io.IOException;
import java.util.Iterator;
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.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.TextOutputFormat;
public class CsvMR {
public static class Map extends MapReduceBase implements
Mapper<LongWritable,Text,Text,IntWritable>{
private final static IntWritable one = new IntWritable(1);
@Override
public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable>
output, Reporter reporter)
throws IOException {
String line = value.toString();
String[] data = line.split(",");
output.collect(new Text(data[2]), one);
}
}
public static class Reduce extends MapReduceBase implements Reducer<Text,
IntWritable, Text, IntWritable>{
@Override
public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text,
IntWritable> output,
Reporter reporter) throws IOException {
Text t_key = key;
int frequency = 0;
while(values.hasNext()) {
IntWritable value = (IntWritable) values.next();
frequency+=value.get();
}
output.collect(t_key, new IntWritable(frequency));
}
}
public static void main(String[] args) throws Exception{
JobConf conf=new JobConf(CsvMR.class);
conf.setJobName("transaction");
conf.setOutputKeyClass(Text.class);
conf.setOutputValueClass(IntWritable.class);
conf.setMapperClass(Map.class);
conf.setCombinerClass(Reduce.class);
conf.setReducerClass(Reduce.class);
conf.setInputFormat(TextInputFormat.class);
conf.setOutputFormat(TextOutputFormat.class);
FileInputFormat.setInputPaths(conf, new Path(args[0]));
FileOutputFormat.setOutputPath(conf, new Path(args[1]));
JobClient.runJob(conf);
}
}
```



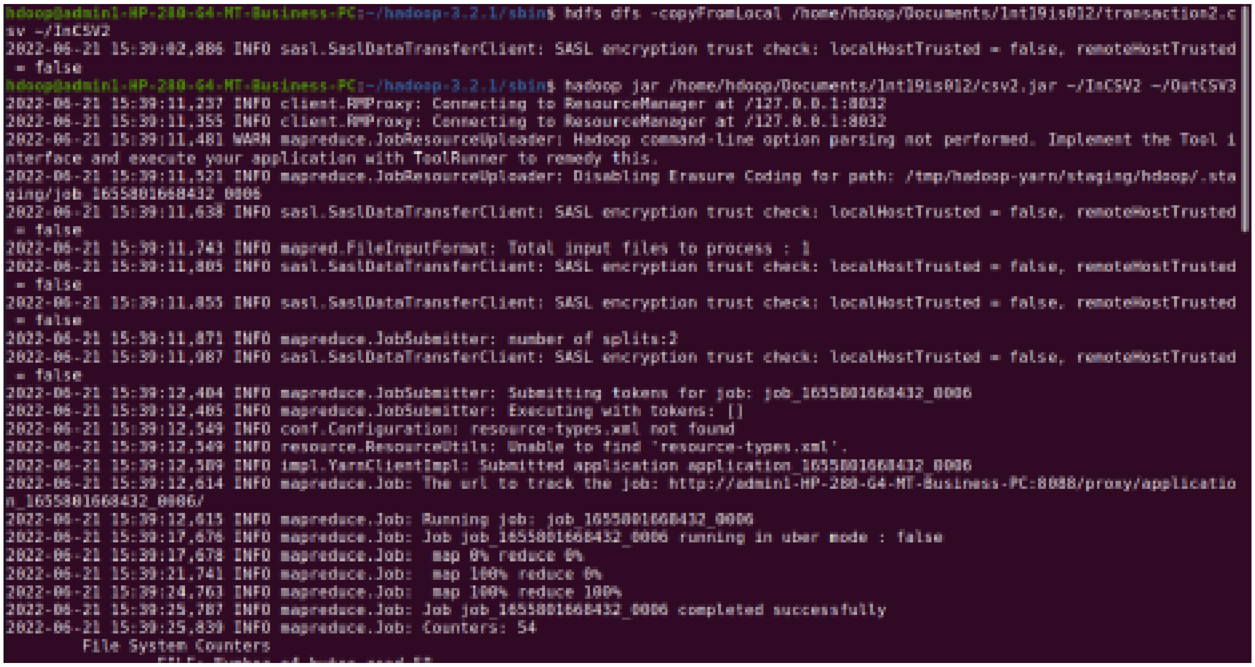
```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -mkdir -p ~/InCSV
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -copyFromLocal /home/hdoop/Documents/1nt19is012/transaction.cs
v ~/InCSV
2022-06-21 15:28:53,859 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hadoop jar /home/hdoop/Documents/1nt19is012/csv.jar ~/InCSV ~/OutCSV
2022-06-21 15:29:23,524 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-21 15:29:23,643 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-21 15:29:23,768 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool i
nterface and execute your application with ToolRunner to remedy this.
2022-06-21 15:29:23,785 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.sta
ging/job_1655801668432_0004
2022-06-21 15:29:23,907 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:29:24,014 INFO mapred.FileInputFormat: Total input files to process : 1
2022-06-21 15:29:24,077 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:29:24,535 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:29:24,551 INFO mapreduce.JobSubmitter: number of splits:2
2022-06-21 15:29:24,667 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:29:24,675 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1655801668432_0004
2022-06-21 15:29:24,675 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-21 15:29:24,788 INFO conf.Configuration: resource-types.xml not found
2022-06-21 15:29:24,788 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-21 15:29:24,829 INFO impl.YarnClientImpl: Submitted application application_1655801668432_0004
2022-06-21 15:29:24,853 INFO mapreduce.Job: The url to track the job: http://admin1-HP-280-G4-MT-Business-PC:8088/proxy/applicatio
n_1655801668432_0004/
2022-06-21 15:29:24,854 INFO mapreduce.Job: Running job: job_1655801668432_0004
2022-06-21 15:29:29,926 INFO mapreduce.Job: Job job_1655801668432_0004 running in uber mode : false
2022-06-21 15:29:29,928 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-21 15:29:33,985 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-21 15:29:37,014 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-21 15:29:38,039 INFO mapreduce.Job: Job job_1655801668432_0004 completed successfully
2022-06-21 15:29:38,094 INFO mapreduce.Job: Counters: 55
```



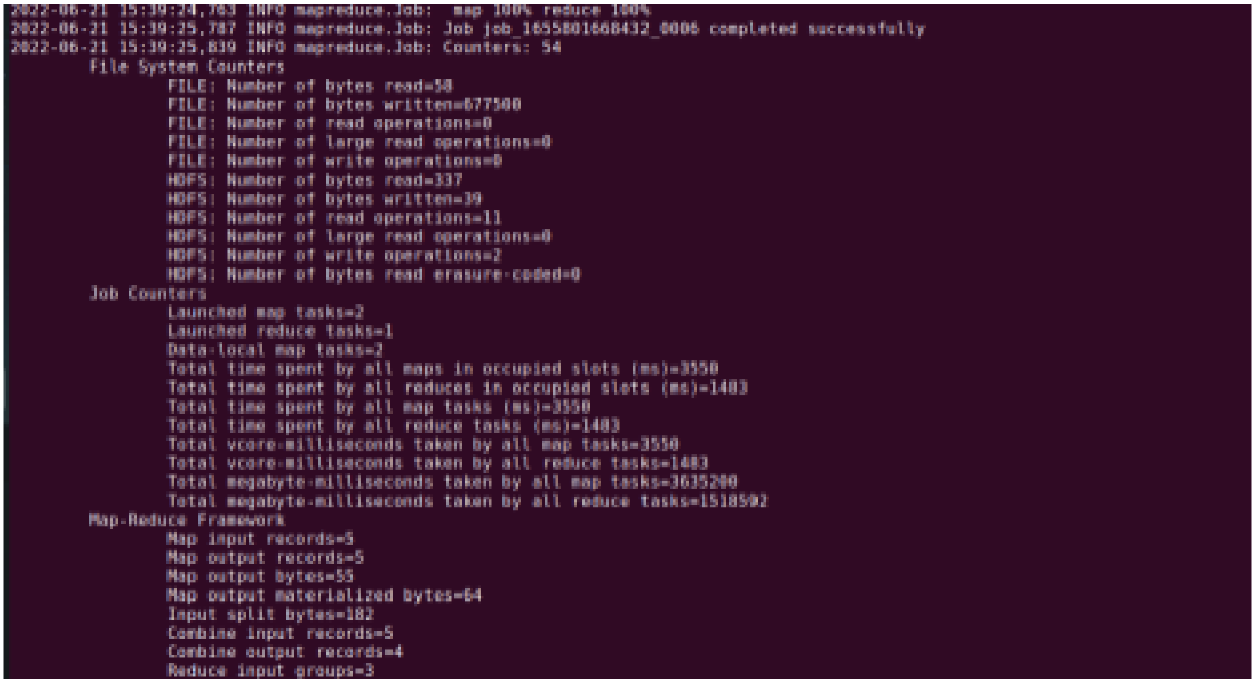
```
                Reduce input groups=4
                Reduce shuffle bytes=66
                Reduce input records=4
                Reduce output records=4
                Spilled Records=8
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=152
                CPU time spent (ms)=1120
                Physical memory (bytes) snapshot=871321600
                Virtual memory (bytes) snapshot=7614525440
                Total committed heap usage (bytes)=876085248
                Peak Map Physical memory (bytes)=335716352
                Peak Map Virtual memory (bytes)=2536763392
                Peak Reduce Physical memory (bytes)=241025024
                Peak Reduce Virtual memory (bytes)=2544676864
        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=203
        File Output Format Counters
                Bytes Written=38
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -cat ~/OutCSV/part*
2022-06-21 15:20:55,680 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
Rajesh  2
Ramesh  1
Soumya  2
Username        1
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ |
```

Part B

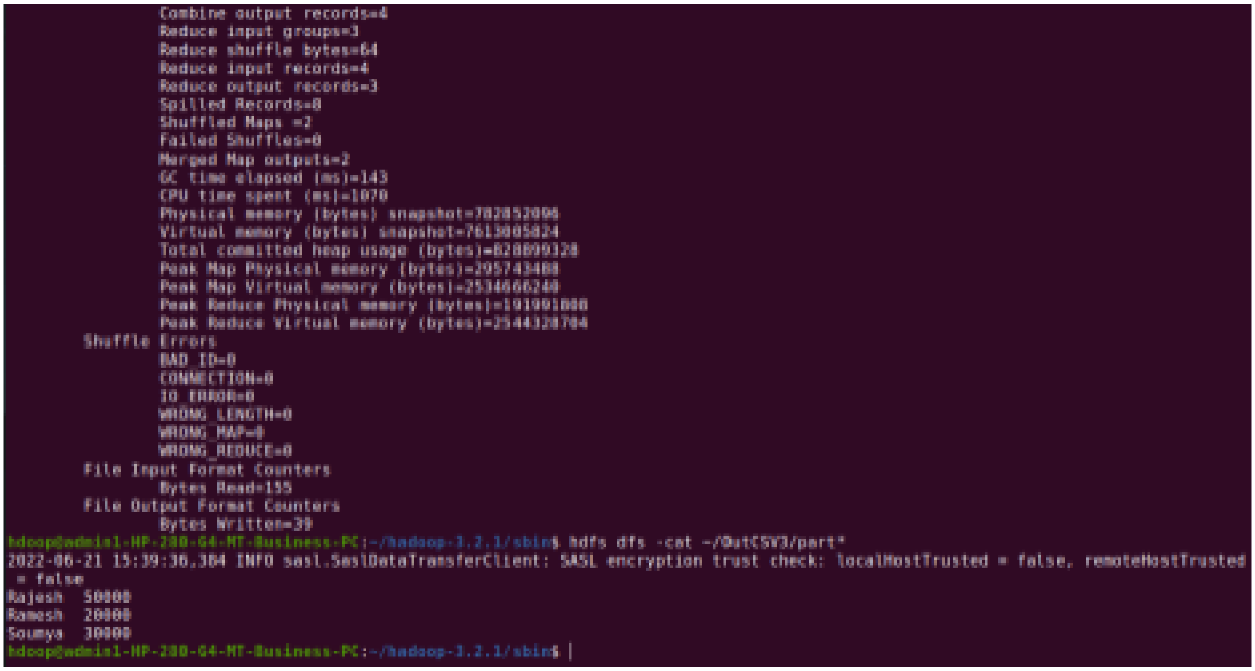
```
package csv;
import java.io.IOException;
import java.util.Iterator;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
import org.apache.hadoop.mapred.TextInputFormat;
import org.apache.hadoop.mapred.TextOutputFormat;
public class CSVMrB {
public static class Map extends MapReduceBase implements
Mapper<LongWritable,Text,Text,IntWritable>{
@Override
public void map(LongWritable key, Text value, OutputCollector<Text, IntWritable>
output, Reporter reporter)
throws IOException {
String line = value.toString();
String[] data = line.split(",");
output.collect(new Text(data[2]), new IntWritable(Integer.parseInt(data[3])));
}
}
public static class Reduce extends MapReduceBase implements Reducer<Text,
IntWritable, Text, IntWritable>{
@Override
public void reduce(Text key, Iterator<IntWritable> values, OutputCollector<Text,
IntWritable> output,
Reporter reporter) throws IOException {
Text t_key = key;
int frequency = 0;
while(values.hasNext()) {
IntWritable value = (IntWritable) values.next();
frequency+=value.get();
}
output.collect(t_key, new IntWritable(frequency));
}
}
public static void main(String[] args) throws Exception{
JobConf conf=new JobConf(CSVMrB.class);
conf.setJobName("transaction2");
conf.setOutputKeyClass(Text.class);
conf.setOutputValueClass(IntWritable.class);
conf.setMapperClass(Map.class);
conf.setCombinerClass(Reduce.class);
conf.setReducerClass(Reduce.class);
conf.setInputFormat(TextInputFormat.class);
conf.setOutputFormat(TextOutputFormat.class);
FileInputFormat.setInputPaths(conf, new Path(args[0]));
FileOutputFormat.setOutputPath(conf, new Path(args[1]));
JobClient.runJob(conf);
}
}
```



```
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -copyFromLocal /home/hdoop/Documents/1nt19is012/transaction2.c
sv ~/InCSV2
2022-06-21 15:39:02,886 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hadoop jar /home/hdoop/Documents/1nt19is012/csv2.jar ~/InCSV2 ~/OutCSV3
2022-06-21 15:39:11,237 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-21 15:39:11,355 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
2022-06-21 15:39:11,481 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool i
nterface and execute your application with ToolRunner to remedy this.
2022-06-21 15:39:11,521 INFO mapreduce.JobResourceUploader: Disabling Erasure Coding for path: /tmp/hadoop-yarn/staging/hdoop/.sta
ging/job_1655801668432_0006
2022-06-21 15:39:11,638 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:39:11,743 INFO mapred.FileInputFormat: Total input files to process : 1
2022-06-21 15:39:11,805 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:39:11,855 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:39:11,871 INFO mapreduce.JobSubmitter: number of splits:2
2022-06-21 15:39:11,987 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
2022-06-21 15:39:12,404 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1655801668432_0006
2022-06-21 15:39:12,405 INFO mapreduce.JobSubmitter: Executing with tokens: []
2022-06-21 15:39:12,549 INFO conf.Configuration: resource-types.xml not found
2022-06-21 15:39:12,549 INFO resource.ResourceUtils: Unable to find 'resource-types.xml'.
2022-06-21 15:39:12,589 INFO impl.YarnClientImpl: Submitted application application_1655801668432_0006
2022-06-21 15:39:12,614 INFO mapreduce.Job: The url to track the job: http://admin1-HP-280-G4-MT-Business-PC:8088/proxy/applicatio
n_1655801668432_0006/
2022-06-21 15:39:12,615 INFO mapreduce.Job: Running job: job_1655801668432_0006
2022-06-21 15:39:17,676 INFO mapreduce.Job: Job job_1655801668432_0006 running in uber mode : false
2022-06-21 15:39:17,678 INFO mapreduce.Job:  map 0% reduce 0%
2022-06-21 15:39:21,741 INFO mapreduce.Job:  map 100% reduce 0%
2022-06-21 15:39:24,763 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-21 15:39:25,787 INFO mapreduce.Job: Job job_1655801668432_0006 completed successfully
2022-06-21 15:39:25,839 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=57
```



```
2022-06-21 15:39:24,763 INFO mapreduce.Job:  map 100% reduce 100%
2022-06-21 15:39:25,787 INFO mapreduce.Job: Job job_1655801668432_0006 completed successfully
2022-06-21 15:39:25,839 INFO mapreduce.Job: Counters: 54
        File System Counters
                FILE: Number of bytes read=58
                FILE: Number of bytes written=677500
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=337
                HDFS: Number of bytes written=39
                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)=3550
                Total time spent by all reduces in occupied slots (ms)=1483
                Total time spent by all map tasks (ms)=3550
                Total time spent by all reduce tasks (ms)=1483
                Total vcore-milliseconds taken by all map tasks=3550
                Total vcore-milliseconds taken by all reduce tasks=1483
                Total megabyte-milliseconds taken by all map tasks=3635200
                Total megabyte-milliseconds taken by all reduce tasks=1518592
        Map-Reduce Framework
                Map input records=5
                Map output records=5
                Map output bytes=55
                Map output materialized bytes=64
                Input split bytes=182
                Combine input records=5
                Combine output records=4
                Reduce input groups=3
```



```
                Combine output records=4
                Reduce input groups=3
                Reduce shuffle bytes=64
                Reduce input records=4
                Reduce output records=3
                Spilled Records=8
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=143
                CPU time spent (ms)=1070
                Physical memory (bytes) snapshot=782852096
                Virtual memory (bytes) snapshot=7613005824
                Total committed heap usage (bytes)=828899328
                Peak Map Physical memory (bytes)=295743488
                Peak Map Virtual memory (bytes)=2534666240
                Peak Reduce Physical memory (bytes)=191991808
                Peak Reduce Virtual memory (bytes)=2544328704
        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=155
        File Output Format Counters
                Bytes Written=39
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ hdfs dfs -cat ~/OutCSV3/part*
2022-06-21 15:39:36,384 INFO sasl.SaslDataTransferClient: SASL encryption trust check: localHostTrusted = false, remoteHostTrusted
 = false
Rajesh  50000
Ramesh  20000
Soumya  30000
hdoop@admin1-HP-280-G4-MT-Business-PC:~/hadoop-3.2.1/sbin$ |
```